# Mutations in the Protein Kinase Superfamily


Department of Molecular Biology - Faculty of Sciences

## Doctoral Thesis

Presented by José María (Txema) González-Izarzugaza Martínez
under the supervision of Prof. Alfonso Valencia Herrera
in Madrid in 2011, November $25^{th}$

# PREFACE

# Acknowledgements

With the perspective of time, I now believe that this Doctoral Thesis has been an exciting journey. This trip has proved to be a puzzling, stimulating and even at times an exasperating venture. Above all, it has undoubtedly been an intense learning process that I believe has instilled in me a considerable degree of scientific maturity, which I would like to think extends to the personal level as well.

It is now time to be humble and while this Thesis may be complete, the journey is not yet over. Several issues remain unsolved and several questions remain unanswered. New thrilling projects will follow, such that this Thesis does not represent the end of the road but rather a junction that may be followed by new beginnings.

In these years, I have had the opportunity to work with a group of brilliant and committed scientists that have inspired me a lot, day by day, step by step, scientifically and personally. So many years, so many people, it is difficult to name you all and I am sure I will miss someone out: please forgive me and consider yourself included.

Alfonso, from you I have learnt what Computer Biology means. From the very beginning, way back when I attended the Masters degree, I found your visionary understanding of this exciting field utterly inspiring. You trusted me and gave me the opportunity to be here today. For that I am very grateful, many thanks.

I want to thank you, Belen, for your patience and sympathy. If Alfonso is the head, you are the heart of the group. This can be extended to Gema also and, previously to Marta too. Many thanks for your daily help.

Many other brilliant scientists deserve to be mentioned here, each of you know why: David - a living Encyclopedia of Science -, Paolo, Raquel, Kris, Jorge, Cesar, Nat, Sito, Ana, Alonso, Antonio, Mark, Angela, Jose Maria, Angel, Edu... In general all the, current and former, members of the Structural Computational Biology Group and the Bioinformatics Unit.

Christine and Andrew, many thanks for hosting me in London, as well as to the rest of laboratory 636 at the UCL who made me feel so welcome: Anja, Lisa, Kathrin, Benoit, Robert, Ian, Corin, Dave, Phil, Jon, Oliver, Ali, Adam, Jim... It was a wonderful experience, personally and scientifically. Many thanks.

Many thanks to my running pals and to my beloved fellow Penguins, particularly to you Michael. A long time has passed since we started this project back in 2006 and it is still afloat. Hockey has provided me with plenty of good things: loyalty, respect, balance... but among them, you - Caro - are definitely the best.

Many thanks also to my friends and family, and I especially want to mention here my Godmother Silvia and my Goddaughter Olaia. Many people accompany me today: from Bilbao, from Castro, from Madrid, from London... Life would be hard without people like you to share the successes and burdens. Again, many thanks.

Finalmente, muchas gracias a mis padres. Siempre me apoyáis y estáis conmigo donde quiera que el camino me lleve. Espero que estéis tan orgullosos de mi como yo lo estoy de vosotros.

Muchas gracias a todos, de verdad.

"*If I have seen further it is only by standing on the shoulders of giants*" - Isaac Newton

# Summary

Protein Kinases constitute a promising pharmaceutical target since they are involved in a large number of tumorigenic functions such as immune evasion, proliferation, anti-apoptosis, metastasis and angiogenesis. Although a small number of single-nucleotide kinase aberrations are causally associated with human diseases, most of the many protein kinase mutations published in the literature are tolerated and therefore, they are neutral in terms of protein and cell activity.

The mechanisms by which mutations elicit aberrant phenotypes have been studied and characterized in some relevant cases for which cause-effect relationships are now well understood. Nevertheless, the biochemical characterization of mutations cannot keep up with the pace of current high-throughput mutation discovery technologies, since the detailed study of each single mutation requires an enormous amount of effort, time and resources.

Thus, there is a clear need to broaden our understanding of mutations in the protein kinase superfamily and, in particular, the mechanisms by which these alter protein function and cause disease. This will help to develop cost-efficient protocols to identify, annotate, characterize and prioritize mutations, such that communal efforts can focus on those most likely to play a direct role in human disease.

The aims of this thesis are to extend our understanding of the mechanisms by which pathogenic mutations disrupt the structure and function of protein kinases, and to design a reliable pipeline to identify mutations likely to be causally implicated in human disease. Focusing our efforts on protein kinases makes possible the use of methods and ideas regarding their specific evolution and organization in protein families. This exclusive information can yield better results than those achieved by general-purpose methods.

Such challenging biological problems will be tackled in this doctoral thesis from the perspective of Computational Biology, a discipline that provides a powerful framework for the integrated analysis of complex information from multiple sources. Current advances in biostatistics and automatic machine learning technologies enable generalized rules to be established based on prior observation, which can then be used to assess the probability of newly discovered mutation being harmful.

# Index of Contents

# DISSERTATION

Chapter 1

# Introduction

## 1.1 Genomic diversity in the cell

### 1.1.1 Defining genomic diversity in the cell

The first draft of the human genome was made publicly available at the beginning of this century [1]. Although it undoubtedly represents a remarkable milestone in scientific research, the existence of a human genome common to all individuals is an unrealistic simplification. Every individual, except identical twins, has its own genome that is only approximately 99.9% identical to that of the rest of the human population. Although this 0.1% might look like an insignificant difference at first glance, it represents more than 3 billion base pairs. This variable fraction in the genome accounts for the different phenotypes in the population and it explains most of the varying susceptibility to diseases [2].

The majority of these differences are single base variations. Indeed, single nucleotide variations are the most common form of genetic diversity in humans, comprising approximately 90% of sequence polymorphisms [3, 4] at a frequency of about 1 mutation per 200-1000 nucleotides [4–7]. These variations include transitions, transversions, and single base insertions and deletions (*indels*). Transitions are base changes from purine to purine or from pyrimidine to pyrimidine, whereas transversions change a purine to a pyrimidine and *vice versa*. Other forms of genetic variation not considered here include microsatellites, longer simple sequence repeats, copy number variants, transposable element insertions, large deletions, chromosome inversions, chromosome translocations and aneuploidy, among others. Even though the analysis of their consequences would also produce interesting results, this is beyond the scope of this thesis.

It is crucial to differentiate three closely related terms that are quite often misused in the literature to refer to single nucleotide divergences: mutation, variation and polymorphism. *Mutation* describes an allele that deviates from the wild type (the majority type), especially if the aberrant allele alters the organisms phenotype or leads to disease. The other two, *variation* and *polymorphism*, are closely related terms with a common origin. Both variations and polymorphisms originally arose as mutations in that the conversion of one nucleotide into another is a purely mutational event. By the time a sequence variant is observed in a given population, this original episode has long passed and the observed divergence is no longer a mutation but rather, a rare sequence variant or a polymorphism depending on the observed frequency. Accordingly, the widely accepted value to differentiate between these two states is a minor allele frequency of at least 1% in the studied population [3, 4]. It is quite common to refer to single nucleotide polymorphisms as SNPs.

However, several paradoxes arise. First, polymorphisms in one population might be rare variations in another. Second, in the genomic era, most SNPs are first detected in a small sample (<10 individuals) and hence, this population frequency criterion cannot be applied robustly. Therefore, the definition of SNP is subtle and most mutations should initially be described as candidate SNPs [8].

### 1.1.2 Sorting out genomic diversity

There are plenty of ways to classify genomic diversity, each relying on different criterion, scope and aims (for a detailed review, refer to [8]). Firstly, we can categorize mutations according to the chemical nature of the bases involved. Previously, we defined *transitions* as changes that do not alter this chemical nature since they involve changes from purine to purine or of pyrimidine to pyrimidine (A to G, C to T, G to A and T to C). By contrast, *transversions* involve changes from a purine to a pyrimidine and *vice versa* (A to C, A to T, G to C, G to T, C to A, C to G, T to A and T to G). Interestingly, although there are twice as many possible transversions than transitions, both occur at similar frequencies, which to some extent might be explained by the genetic code. Indeed, transitions are less likely to affect amino acid sequences than transversions and thus, they are thought to have a higher probability of retention in coding regions [8].

Based on their genomic location, polymorphisms can be classified as *coding and non-coding*. Non-coding mutations are located in the introns of a gene locus, intragenic regions, transcription factors binding domains and any other genomic region that does not yield a protein product. Coding mutations occur in DNA regions that lead to a protein product, that is, in the exons of the coding sequence.

Additionally, coding polymorphisms can be further sub-classified depending on whether they alter the resulting protein product or not. *Non-synonymous* variations alter the amino acid sequence of the protein through either amino acid substitution (missense polymorphisms) or the insertion of truncation mutations (nonsense polymorphisms). By contrast, *synonymous* variations (also described as silent) are those that do not alter the amino acid sequence of the protein even if the sequence is altered at the genomic level due to the degeneracy of the genetic code.

We can also classify mutations according to their origin as *germline* or *somatic* [9] (Figure 1.1). If the mutation is inherited and, consequently, it is present in every cell of the individual, we define it as germline. Conversely, if the mutation is acquired after conception, for instance through the effects of carcinogens, UV light or problems during DNA replication, we call it somatic. These mutations are not passed on to the offspring and they are not present in all the cells of the individual. Consequently, if the mutated cell continues to divide, it generates a patch of tissue expressing a genotype that differs from that of the rest of the body. Cancer is a good example of such events and we will discuss the implication of somatic mutations in current models of cancer onset and progression later in this chapter.

### 1.1.3 Genotyping studies to detect genes harboring disease-associated mutations

Novel point mutations that affect gene function have historically been identified by small-scale sequencing of individual (or a few) genes. Technical advances in the last decade have favored

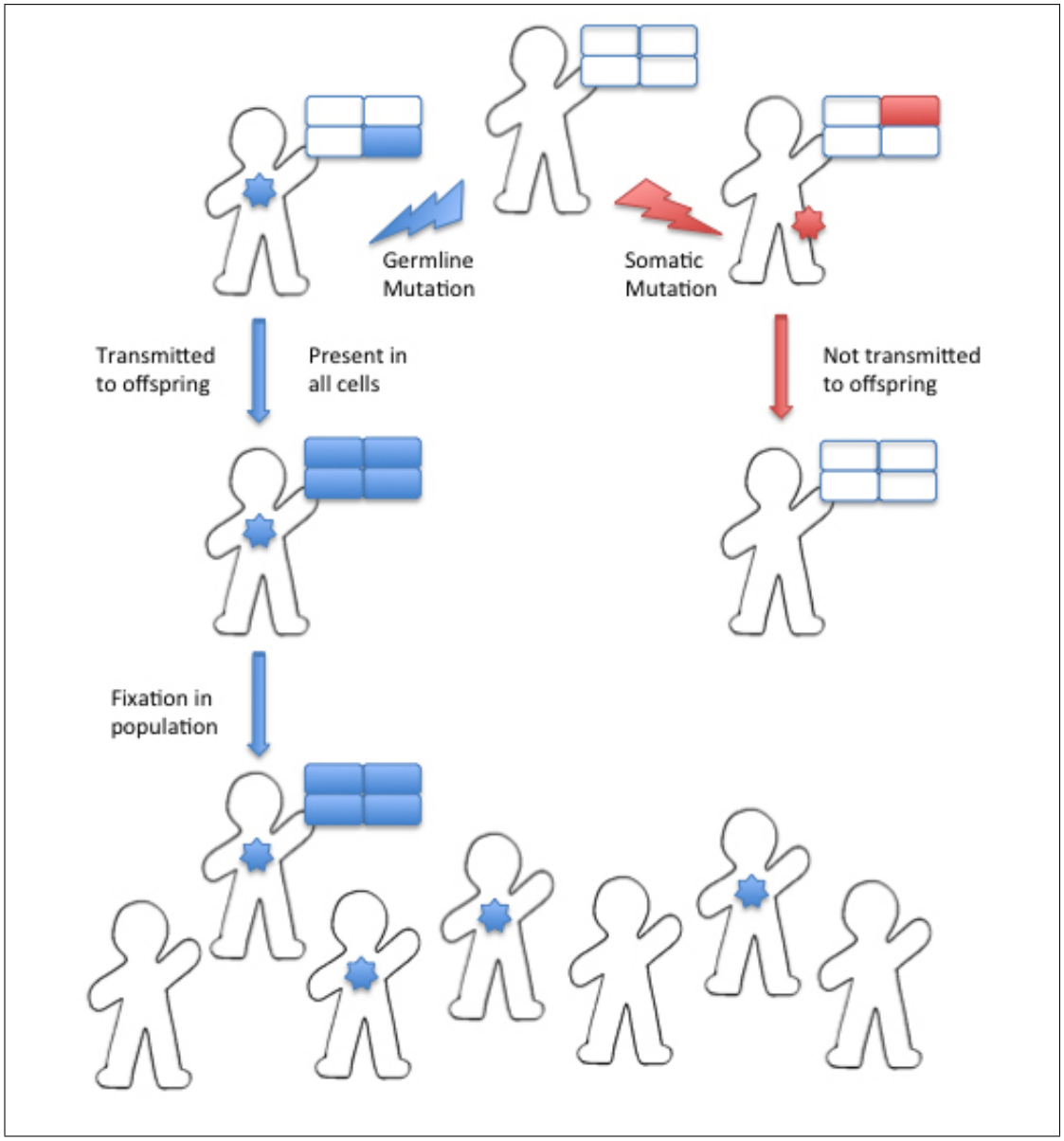



**Figure 1.1:** Differences between somatic and germline mutations. Germline mutations (blue star) are transmitted to the offspring and they are present in all the cells of the offspring (profile of filled squares). If a mutation becomes fixed in the population, it becomes a polymorphism. Conversely, somatic mutations (red star) are not present in all the cells of the individual and they are not inherited.

the development of high-throughput methods to detecting candidate disease-associated genes. High-throughput resequencing screening can detect polymorphisms as well as small 'indels' [10], and the introduction of 'next generation sequencing' (NGS) technologies [11] has not only produced massive amounts of data, but it has also permitted the quantitative identification of individual gene variants and the detection of abnormal transcripts [12].

Current methods to identify mutated candidate genes follow two closely related yet different approaches, as reviewed recently [13]. On one hand, a number of methods are based

on sequencing a relatively small subset of genes (for instance, Greenman [14] analyzed the genes encoding the human kinome), which are then genotyped in a large number of samples. Although this approach may identify candidate genes harboring mutations at low frequencies, the major drawback is that it requires *a priori* gene selection (also referred to as *prioritization*). On the other hand, there are methods that analyze the coding sequences of whole genomes in a smaller number of samples, such as colon and breast tumors [15, 16], pancreas adenocarcinomas [17] or glioblastomas [18]. This approach identifies the genes most frequently mutated in the context of disease.

The most important high-throughput cancer genomic projects from the last decade have been summarized in Table 1.1 (extracted from our 2009 revision, Baudot *et al.*, 2009 [13]) and as can be seen, protein kinases and cancer samples are very amenable targets for the approaches discussed here (the relationship between cancer and protein kinases will be studied more thoroughly below). Indeed, cancer is an interesting example of a complex disease with a profound genomic component.

| **Author** | **Year** | **Genes** | **Tumors** | **Screening size** |
|---|---|---|---|---|
| Bardelli | 2003 | Tyrosine kinase | Colon | 138 genes, 35 samples, a subset in 147 additional samples |
| Wang | 2004 | Tyrosine phosphatase | Colon | 87 genes, 18 samples, a subset in 157 additional samples |
| Stephens | 2005 | Kinase | Breast | 518 genes, 25 samples, a subset in 56 additional samples |
| Davies | 2005 | Kinase | Lung | 518 genes, 33 samples, a subset in 56 additional samples |
| Sjblom | 2006 | All | Breast / colon | 13023 genes, 22 samples, a subset in 48 additional samples |
| Greenman | 2007 | Kinase | Several | 518 genes in 210 human cancers |
| Wood | 2007 | All | Breast / colon | 18191 genes, 22 samples, a subset in 48 additional samples |
| Loriaux | 2008 | Tyrosine kinase | Acute myeloid leukemia | 85 genes, 188 samples |
| Tomasson | 2008 | Tyrosine kinase | Acute myeloid leukemia | 26 genes, 94 samples, a subset in 94 additional samples |
| Brown | 2008 | Tyrosine kinase | Chronic lymphocytic leukemia | 70 genes, 95 samples |
| Jones | 2008 | All | Pancreas | 20661 genes, 24 samples |
| Parsons | 2008 | All | Glioblastoma | 20661 genes, 22 samples, a subset in 83 additional samples |
| CGARN* | 2008 | Custom | Glioblastoma | 601 genes, 91 samples |
| Ding | 2008 | Custom | Lung | 623 genes, 188 samples |

**Table 1.1:** Catalogue of the main recent high-throughput cancer genomic studies and initiatives. Adapted from Baudot *et al. (2009)* [13]. CGARN: Cancer Genome Atlas Research Network.

### 1.1.4 The 'driver gene, passenger gene' metaphor

During a keynote lecture in 1964, Sir Christopher Andrewes coined the terms drivers and passengers to refer metaphorically to the role of certain viruses in either causing cancer or merely being passengers in infected cells [19]. Nowadays, the terms have slightly different meanings, but the underlying metaphor is still valid. The term *driver* is often used for somatic mutations that are positively selected and contribute to tumor development or progression, whereas the term *passenger* is used for cancer-neutral mutations that are retained during the evolution of cancerous cells but do not affect tumor progression [9]. This classification can be extended to the genes that harbor these mutations as well.

To identify the noxious mutations involved in cancer it would clearly be logical to isolate these driver genes. An interesting predictive strategy would be to use the overall frequency of mutations in a gene to detect the positive selection that could explain its behavior as a driver of oncogenesis. The assumption here is that positive selection is exerted on non-synonymous mutations. The degeneration of the genetic code yields a 2:1 ratio of non-synonymous mutations to synonymous ones and thus, higher ratios indicate positive selection due to a competitive advantage. In reality, more complex models borrowed from the field of molecular biology apply, these taking into account the types of mutation (transition or transversion), the sequence context (for example a C to A transversion in a CpG island) and the conservation of the amino acids associated with the mutation [14, 20–22]. Another group of similar methods account for the differences between the observed frequencies of non-synonymous mutations and the estimated expected frequencies. These approaches are based on the rationale that the presence of genes mutated in a number of tumors would suggest that positive selection has occurred during tumorigenesis, which could be taken as an indicator of the selective advantage conferred by that gene during the process of tumorigenesis. The frequency expected for the background mutations is commonly derived from the synonymous mutation rates, which are assumed to be independent of selection. To the best of our knowledge, the first publication applying this kind of statistics to predict driver genes was published in 2006 [15]. In this study, the probability of a gene containing a given number of mutations was computed according to gene size. However, this approach did not consider that the background mutation and repair rates vary between tissues, chromosomes, regions and genes. A later study [16] complemented this analysis by including chromosome-specific background mutation rates, correcting the ratio of non-synonymous:synonymous mutations and investigating the potential effects on protein function.

The obvious limitation of these approaches to predict driver genes, apart from the need of a large number of samples to provide statistical significance, is that they do not provide information about the specific alleles (point mutations) but only the genes harboring them. This is an important problem, since from the vast amount of mutations that are detected by current high-throughput approaches, only a handful are directly implicated in disease, while the others are neutral [13, 14]. The consequences of individual mutations in disease will be discussed in the following sections.

### 1.1.5 Mutations and disease

Disease is a harmful kind of altered phenotype. Since the early days of genomics, association studies have demonstrated the relationship between an aberrant allele and its pathogenic manifestation, with experimental validation of the biochemical effects of a given alteration considered as proof of a mechanistic involvement. Indeed, later in this chapter, we will introduce some examples of key mutations within the human kinome causally involved in

certain types of cancer and in drug resistance.

The 1000 Genomes project [4] (`www.1000genomes.org`) aims to characterize the human variome from a broader perspective than the Human Genome project [1]. Thus, it is devoted to the genome-wide high-throughput sequencing of a broad set of individuals from different populations. In fact, the project includes three sub-projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother-father-child trios; and exon-targeted sequencing of 697 individuals from seven populations. Despite being at a very early stage, the results that have been obtained are striking: 15 million single nucleotide polymorphisms, in addition to 1 million short insertions and 20,000 structural variants, have been characterized. Most of them are neutral and are not causally associated with any aberrant phenotype. Nevertheless, on average, each person has been found to have around 250 to 300 variations that disrupt protein function and 50 to 100 variants previously implicated in inherited disorders [4].

### 1.1.6 Somatic mutations in cancer: cancer evolution models

Cancer is a complex disease of which certain aspects are still not fully understood [13, 23–26]. Over the last couple of decades, a number of models have attempted to shed light on diverse aspects of the disease, such as the different contribution of mutations to tumor development and progression [9, 13, 27].

Historically, the prevailing model has been the *clonal evolution theory* (Figure 1.2, blue pathway), where cancer cells accumulate specific genetic changes that lead to clonal expansion. These changes are selected through a competitive Darwinian process, where the mutations conferring a selective advantage become fixed, thereby permitting phylogenetic tracing of the history of the evolving cancer cell populations. Thus, it is generally accepted that malignant tumors arise from benign precursors, whereby genomically stable tumors precede their genomically unstable counterparts, and that metastasis is the final step in tumor evolution.

Unfortunately, not all results fit this simplified model. For instance, some breast cancer metastases do not resemble the primary tumor [28], while a small proportion of normal mouse mammary epithelial cells injected intravenously can survive at distant sites and eventually produce tumors [29]. These observations have generated a new model, the *parallel evolution* (Figure 1.2, red pathway), where cells that generate metastases are segregated relatively early from the primary tumor and evolve independently [30, 31]. Gene-expression profiles that predict metastases in certain primary breast tumors support this model [32].

The differences between these two models affects the interpretation and clinical implementation of the data gleaned about cancer-associated mutations. In terms of the clonal evolution theory, the relevance of a mutation detected in a secondary (metastatic) tissue is unclear in the absence of information about its presence in primary tumors. Moreover, in the parallel evolution model, targeting this mutation for therapeutic treatment would have little effect on treating the primary tumor, which would evolve independently.

Although cancer is a commonly understood as a single disease, this term more likely refers to a growing group of more than 100 closely related complex diseases with diverse risk factors and epidemiology [9]. Given the complexity of cancer and the diversity of its phenotypic outcome, it is very unlikely that a single model can universally cover all aspects of these diseases. However, it is plausible that both models are complementary (Figure 1.2, green pathway) and such an integrated approach is supported by a recent study indicating that metastases could act as repositories from which systemic tumor cell settlements could occur [33].

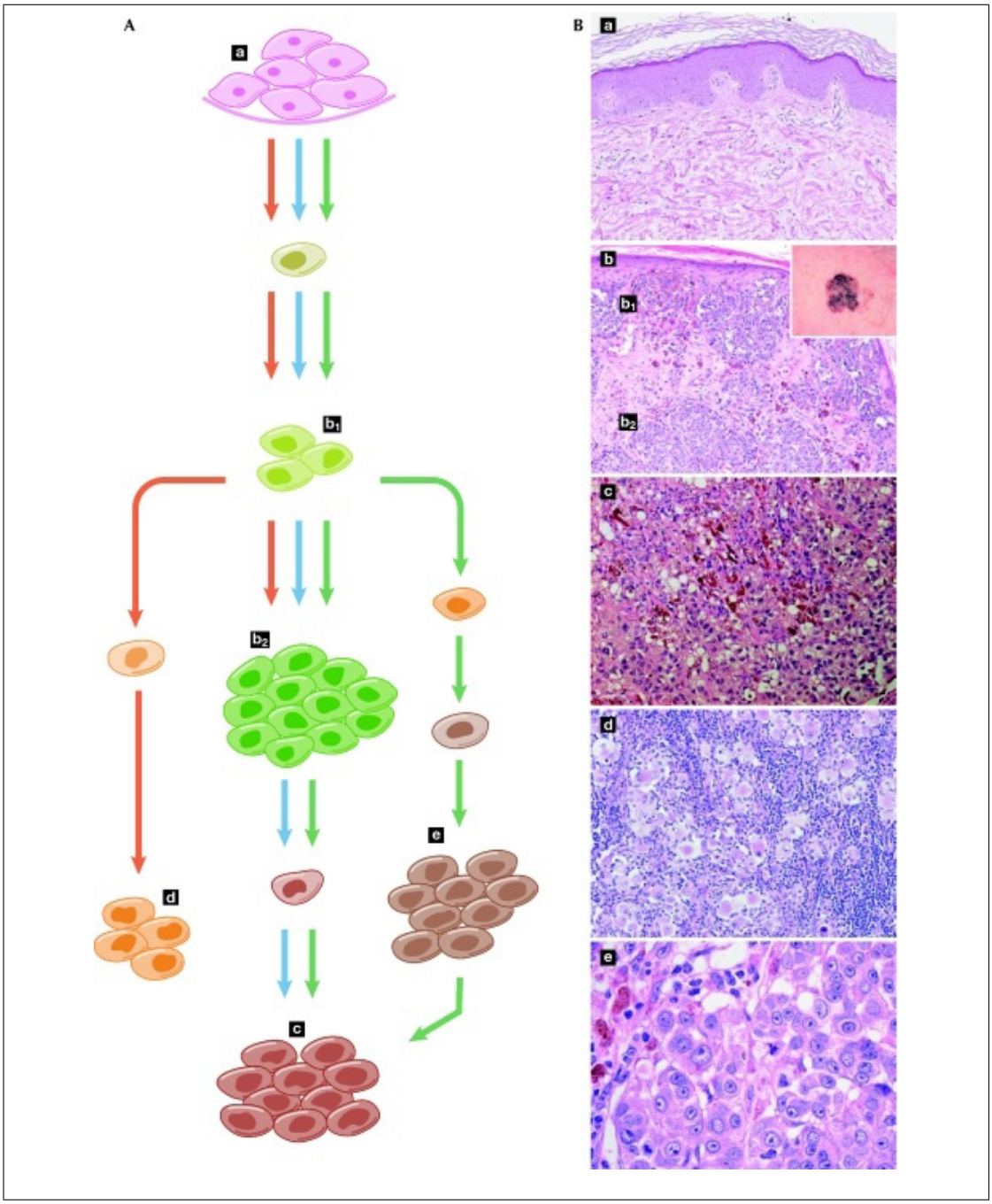



**Figure 1.2:** Models of cancer evolution. The *clonal evolution* theory (blue arrows) is the prevailing view to explain the successive steps of mutation and selection from normal tissue to a primary tumor and metastasis. However, metastasis-generating cells can emerge relatively early in tumorigenesis and seed distant tissues, thereby evolving in parallel with the primary tumor and defining the *parallel evolution* model (red arrows). These two models can occur simultaneously, and metastatic deposits may serve as sites from which additional metastases can be generated, thereby leading to an *integrated* model of cancer evolution (green arrows). The microphotographs illustrate the paradigms modeled (taken from Baudot *et al.* (2009) [13]).

### 1.1.7 Mutations, protein kinases and cancer

Although about 1% of all human genes are known to contribute to cancer through acquired mutations, the family of genes most frequently contributing to cancer is the protein kinase superfamily [34]. Protein kinases are already a promising pharmaceutical target given that they are involved in a large number of tumorigenic functions such as immune evasion, proliferation, anti-apoptosis, metastasis and angiogenesis, possibly due to the simplicity of the mechanism of attaching an ATP-derived phosphate to a substrate protein [35]. In fact, a growing number of clinical drugs already target members of the protein kinase superfamily [36, 37], the first of which was Fasudil, approved back in 1999. Originally developed to treat cerebral vasospasm, Fasudil is a powerful competitive inhibitor of Rho kinase that binds to the ATP-binding site. Nowadays, the most powerful compound to treat cancer is Imatinib (Glivec), an inhibitor of the Abl tyrosine kinase used to treat Philadelphia chromosome-positive chronic myeloid leukemia. Since the disease arises from a single mutation that leads to an aberrant Abl kinase, this drug is fully effective if the disease is discovered at an early stage. Interestingly, although this inhibitor binds in a competitive manner to the ATP-binding pocket, its binding extends to a region behind the C-helix of the N-terminal lobe. However, proper binding can only be achieved in the inactive conformation of the kinase. For this reason, at an advanced stage of the disease when more mutations appear in the protein, the drug is not so effective. In particular, a threonine to leucine mutation at a position next to the C-helix pocket prevents Glivec binding and thus, it enables escape from the protective action of the drug. This is not the only example where mutations affect protein kinase function and indeed, there is now a plethora of kinase mutations that are thought to be important for cancer onset and development. They will be discussed later in this doctoral thesis.

Another very interesting example of the role of kinases in cancer involves the cyclin-dependent kinase (CDK). Progression through the cell cycle is screened by several checkpoints that detect defects during chromosomal replication and segregation. If any of these control points detects an error, the cell cycle is arrested, which enables the cell to attempt to repair any defects found and prevent their transmission to its offspring. This arrest is mainly achieved by modulating CDK activity (Figure 1.3).

Due to the crucial role of CDKs, their activity is intricately regulated. CDK activity requires the binding of regulatory subunits known as cyclins, which are synthesized and destroyed at specific times during the cell cycle to enable the CDKs to act so precisely. There are three interphase CDKs (CDK2, CDK4 and CDK6) and a mitotic CDK (CDC2/CDK1), and in addition, there are ten different cyclins grouped into four different classes (A, B, D and E). Several mutations have been reported to de-regulate certain CDK-cyclin complexes, which result in either continued cell proliferation or unscheduled re-entry into the cell cycle, both characteristics typical of tumorigenic cells [26].

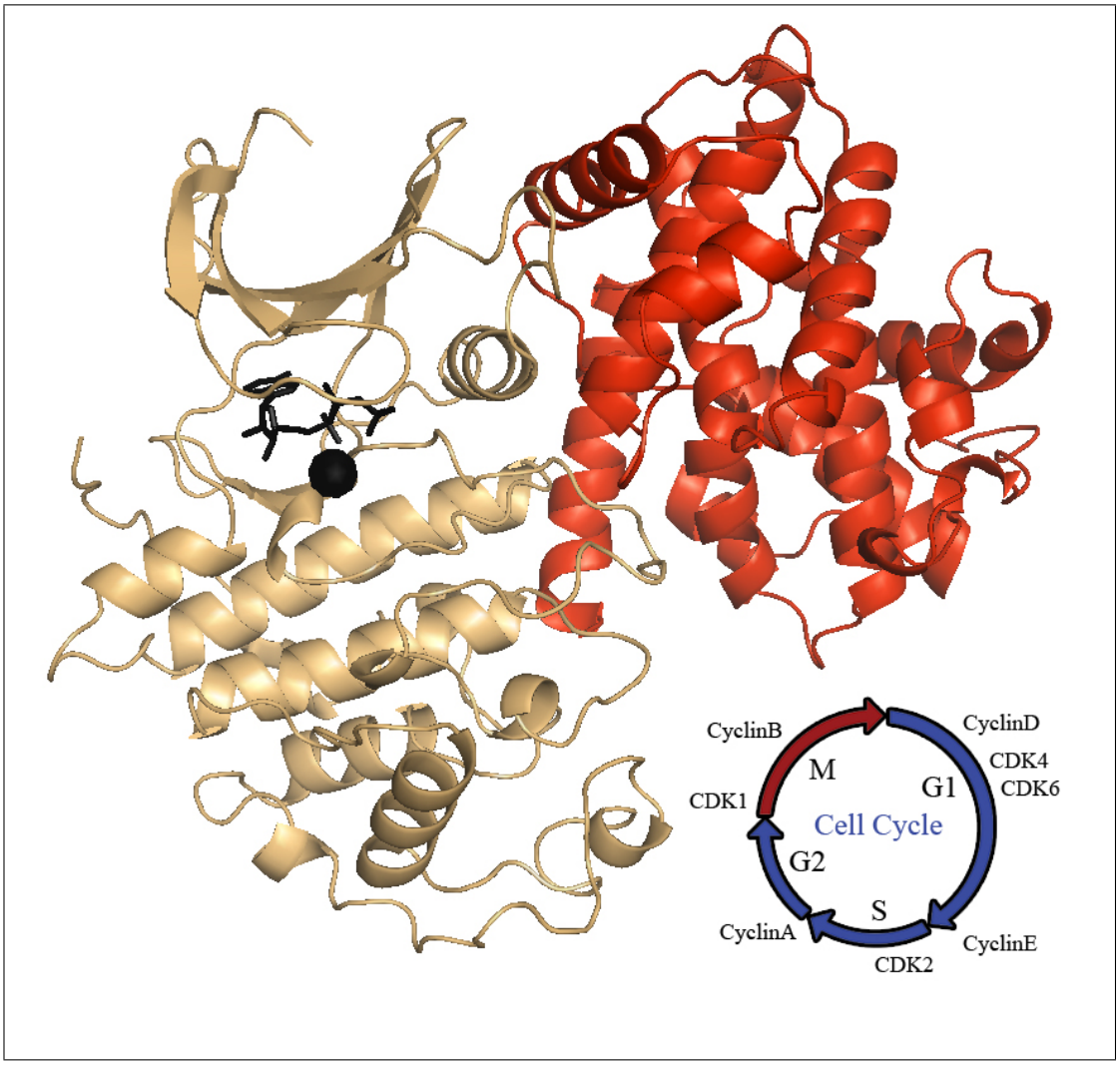



**Figure 1.3:** CDK2 interaction with Cyclin A regulates the cell cycle. Active human CDK2 (orange) in complex with Cyclin A (red). Time-regulated differential expression of cyclins and CDKs regulates the cell cycle.

## 1.2 The human kinome

### 1.2.1 Metabolic switches in the cell: protein kinases and phosphatases

*"The conversion of phosphorylase b to a, as it occurs in cell-free muscle extracts, requires a nucleotide containing high energy phosphate in addition to a divalent metal ion. Whether this implies that during conversion there is a direct phosphorylation of the enzyme or the formation of an active intermediate cannot be stated at this time."* Edmond H. Fischer and Edwin G. Krebs stated this in 1955 when, for the first time, they introduced the concept of reversible phosphorylation as a plausible mechanism to regulate protein function [38]. The event they observed was the reaction in which glycogen phosphorylase was phosphorylated by phosphorylase kinase. Phosphorylation switches the inactive glycogen phosphorylase to an active state that is capable of catalyzing glycogen metabolism to provide the energy required for muscle contraction.

It is a well-established dogma nowadays that reversible phosphorylation of proteins is a versatile and recurrent mechanism of biological regulation. Furthermore, all protein families are to some extent affected, directly or indirectly, by phosphorylation. The activity of proteins with very diverse functions is controlled by reversible phosphorylation, from enzymes and transcription factors to scaffolding proteins.

The phosphorylation mechanism is currently well understood. Protein kinases are a family of enzymes that catalyze the transfer of a phosphate from ATP to a serine, threonine or tyrosine hydroxyl group in the target protein. The change to a phosphorylated protein has very different consequences of which activation or inhibition of an enzyme, alteration of interaction surfaces, and conformational changes are among the most common. Due to the importance of the processes regulated, protein kinases generally do not act alone but rather, they form part of a finely tuned signaling cascade that is strictly controlled spatiotemporally. After a signal has been transduced, the phosphate group is then required to be removed to yield an inactive state. This action is performed by the counterparts of kinases, the phosphatases. Therefore, both protein families are metaphorically referred to as the metabolic switches of the cell.

Although both superfamilies deserve detailed study and would provide very exciting observations, we will only focus on the protein kinase superfamily in this doctoral thesis. The main reason for our choice is the growing number of studies that have thoroughly explored the protein kinase superfamily and thus, the large amount of information available for this superfamily. This is particularly relevant if we consider the ever increasing efforts to link mutations in the protein kinase superfamily with cancer.

### 1.2.2 The human kinome

Protein kinases are one of the most ubiquitous families of signaling molecules in the human cell, constituting approximately 2% of the proteins encoded by the human genome. The total number of genes encoding kinases has been a matter of discussion in the last decade and for instance, in 1998, Wang's group predicted there are between 1000 and 2000 different kinases in humans [39]. With the completion of the human genome, the current estimate is that 518 human genes encode protein kinases, corresponding to 1.7% of the total number of genes in the human genome [40].

There are several different classifications of kinases from the main model organisms: yeast [41], worm [42], fruit fly [43] and mouse [44]. The reference classification in humans is KinBase [40], which has also been incorporated into UniProt [45], albeit with minor

modifications.

According to KinBase, there are two different kinase superfamilies. A large superfamily composed of 478 proteins with a eukaryotic protein kinase (ePK) domain and another superfamily containing 40 atypical kinases that lack the conserved domain but that show kinase activity. Additionally, eukaryotic protein kinases can be classified hierarchically into groups, families and subfamilies, mainly based on sequence comparison of the catalytic domain but also, according to the sequence similarity over the whole protein, the domain structure outside the main domain, their known biological function, and other factors including similar classifications in other model eukaryotic families such as yeast, worm and fruit fly. The phylogenetic tree representing the 478 ePK genes and the fraction of kinases in each of the 8 major groups is shown in Figure 1.4.

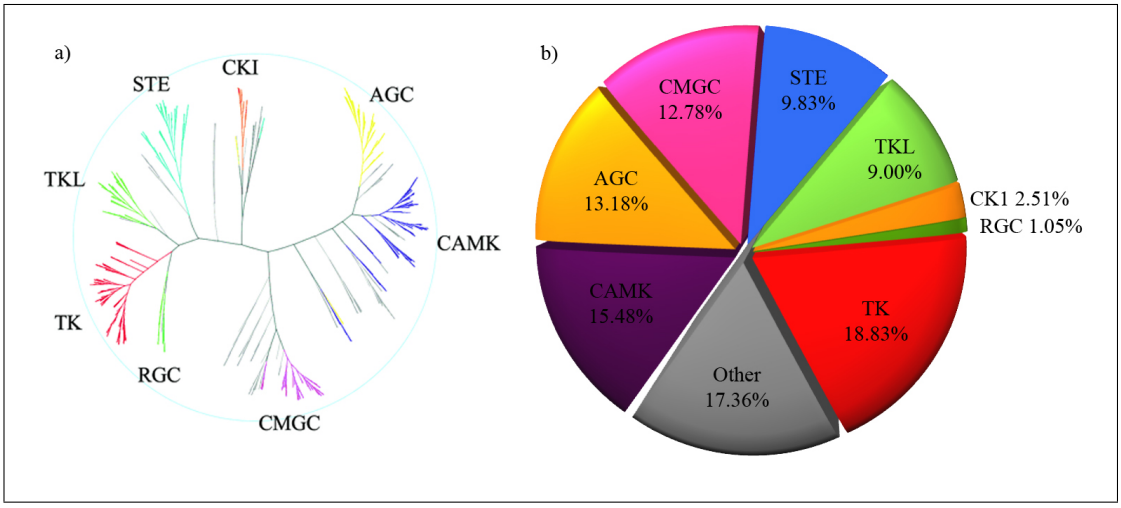



**Figure 1.4:** Classification of the human kinome according to KinBase [40]
Panel A: Dendrogram of eukaryotic protein kinases as described in KinBase.
Panel B: Distribution of human eukaryotic protein kinases in the different KinBase groups.

The distribution of kinases in the different groups for human and commonly used animal models is shown in Table 1.2. Of the 189 different human subfamilies, 51 are present in other model organisms (namely yeast, worm and fruit fly), indicating their functions are important for the eukaryotic cell. Moreover, 93 additional subfamilies are present in at least one of these model organisms but absent in human, suggesting that they evolved to fulfill distinct functions in metazoan evolution. It is interesting to highlight that more than 95% of the human kinases have direct orthologs in mouse [40]. The complete dendrogram of the human protein kinase superfamily is depicted in Figure 1.5.

### 1.2.3 A common feature of all protein kinases: the PK domain

Protein kinases share a common phylogenetic origin that is still evident in their common structural core, the protein kinase (PK) domain. This common framework is flexible enough to accommodate a considerable degree of specific structural variants, as well as a wide range of sequence divergence and substrate specificities, while preserving the conserved basic catalytic mechanism. Indeed, with a few exceptions, kinases share the molecular substructure for binding the ATP-divalent cation complex, and the mechanism to transfer the terminal phosphate of

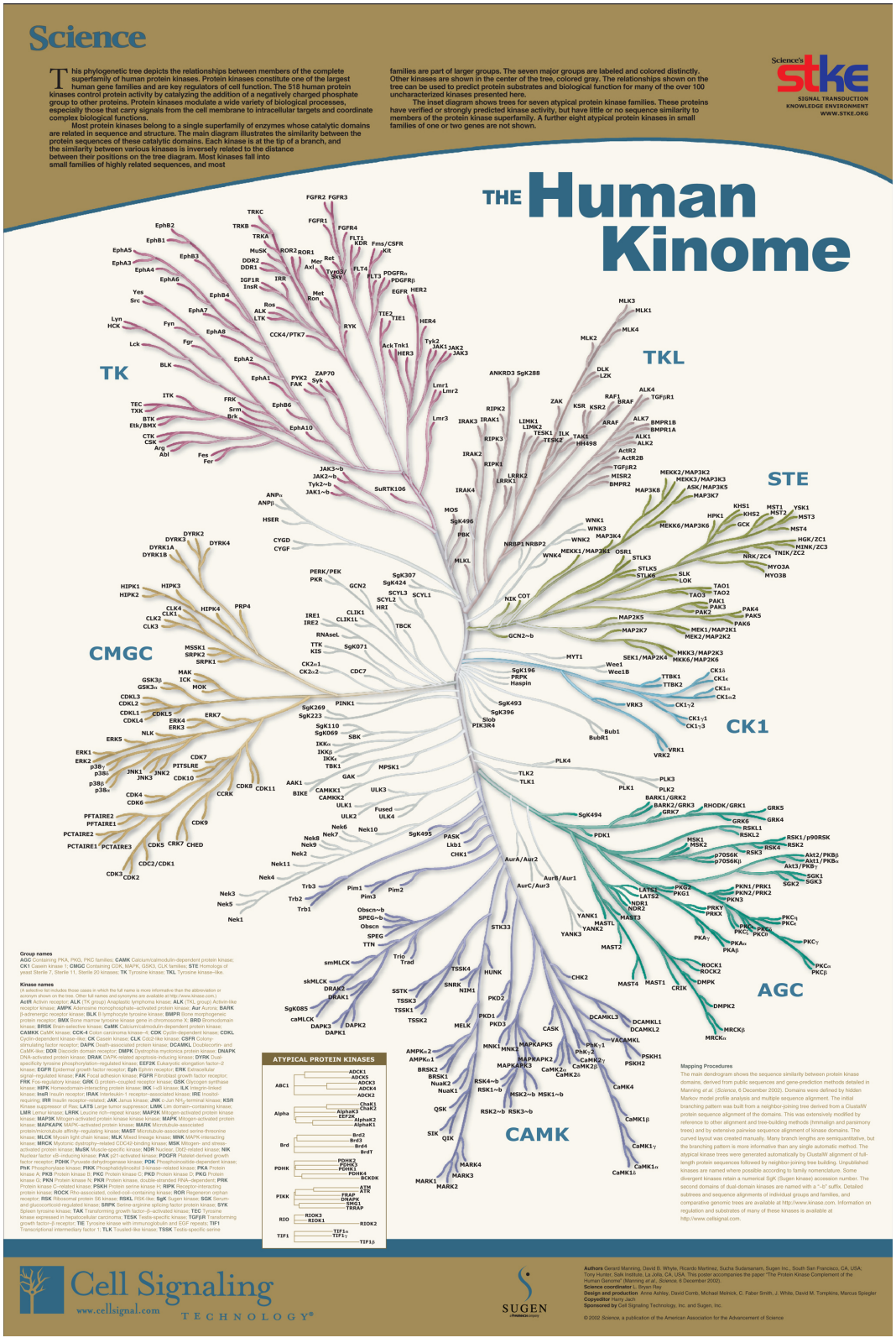



**Figure 1.5:** The Human Kinome. The protein kinase superfamily as published in 2002 [40]

| Class | Group | Fams. | Subfams. | Yeast Kinases | Worm Kinases | FruitFly Kinases | Human Kinases |
|---|---|---|---|---|---|---|---|
| ePKs | AGC | 14 | 21 | 17 | 30 | 30 | 63 |
| | CAMK | 17 | 33 | 21 | 46 | 32 | 74 |
| | CK1 | 3 | 5 | 4 | 85 | 10 | 12 |
| | CMGC | 8 | 24 | 21 | 49 | 33 | 61 |
| | RGC | 1 | 1 | 0 | 27 | 6 | 5 |
| | STE | 3 | 13 | 14 | 25 | 18 | 47 |
| | TK | 30 | 30 | 0 | 90 | 32 | 90 |
| | TKL | 7 | 13 | 0 | 15 | 17 | 43 |
| | Other | 37 | 39 | 38 | 67 | 45 | 83 |
| aPKs | PDHK | 1 | 1 | 2 | 1 | 1 | 5 |
| | Alpha | 1 | 2 | 0 | 4 | 1 | 6 |
| | RIO | 1 | 3 | 2 | 3 | 3 | 3 |
| | A6 | 1 | 1 | 1 | 2 | 1 | 2 |
| | ABC1 | 1 | 1 | 3 | 3 | 3 | 5 |
| | BRD | 1 | 1 | 0 | 1 | 1 | 4 |
| | PIKK | 1 | 6 | 5 | 5 | 5 | 6 |
| | Other | 7 | 7 | 2 | 1 | 2 | 9 |
| TOTAL | | 134 | 201 | 130 | 454 | 240 | 518 |

**Table 1.2:** Distribution of kinases in human and model systems (adapted from Manning *et al.* (2002)). ePKs: eukaryotic protein kinases, aPKs: atypical protein kinases. Fams. and Subfams. represent the number of families and subfamilies in each of the groups.

the ATP to a serine, threonine or tyrosine residue in the target protein [46].

The structure of a canonical protein kinase (PK) domain has a basic two-lobe fold, the N-terminal and C-terminal lobes, joined by a small hinge region. The N-terminal lobe is a set of $\beta$-sheets plus a scaffolding $\alpha$-helix (the conserved $\alpha$C-helix), whereas the C-terminal lobe is mainly a cluster of $\alpha$-helices. $Mn^{2+}$ and ATP bind to a conserved site between the two lobes, and the substrate protein is recognized by interacting with the activation segment in the C-terminal lobe.

Figure 1.6 highlights the most relevant structural elements common to all canonical kinases, represented on the structure of the active conformation of human cyclin-dependent kinase type II (CDK2). In addition to the basic two-lobe structure, there are other structural elements of interest that are often present.

#### 1.2.3.1 The P-loop

This is a small loop at the beginning of the N-terminal lobe that it is very often preceded by a $\beta$-sheet and followed by an $\alpha$-helix. It contains a glycine-rich motif (GxGxxG) that interacts not only with the ATP but also, with the $Mn^{2+}$ ion that coordinates the $\beta$- and $\gamma$- phosphates in the ATP molecule. P-loop stands for phosphate-binding loop.

#### 1.2.3.2 The $\alpha$C-helix

This long helix in the N-terminal lobe acts as a backbone of this subdomain and has a role in ATP binding [46]. In CDKs, this helix contains the evolutionarily conserved PSTAIRE motif, which distinguishes them from other protein kinases and is thought to be involved in the interaction with cyclins.

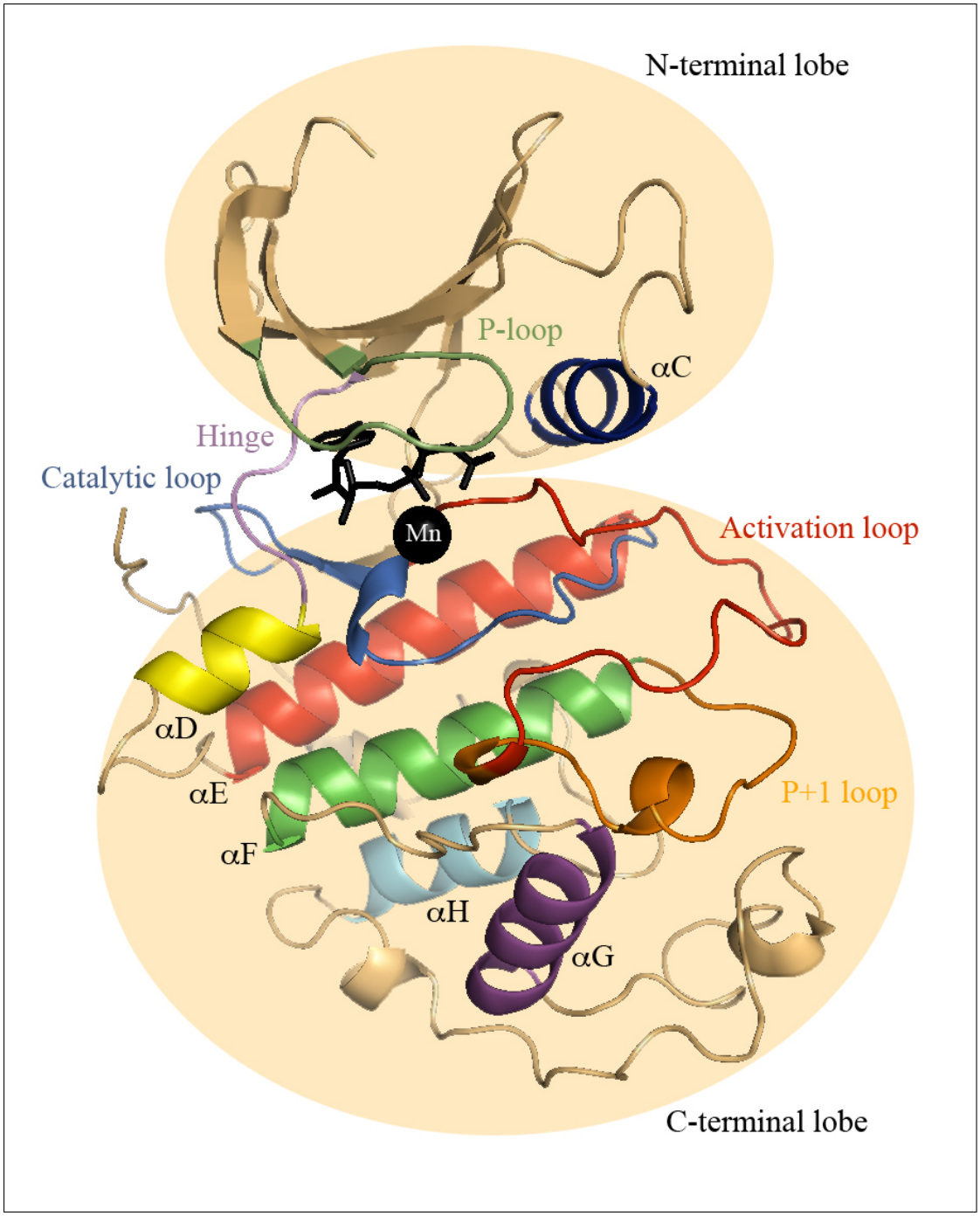



**Figure 1.6:** Active conformation of human cyclin-dependent kinase type II (CDK2). Relevant structural features have been highlighted.

#### 1.2.3.3 The $\alpha$F-helix

Similar to the C-helix, this helix acts as a backbone to the C-terminal domain and it is essential for the assembly of the active conformation of the kinase [47].

#### 1.2.3.4 The catalytic loop

This functionally important loop facilitates the binding and stabilization of ATP.

#### 1.2.3.5 The activation loop (T-loop)

All kinases have a conserved activation loop involved in regulating kinase activity since it can assume a large number of conformations depending on the phosphorylation state, which triggers the active/inactive forms of the kinase. In the active form, the activation loop has an important role in the conformation of the ATP-binding pocket. By contrast, the inactive conformation of the loop blocks the binding of the substrate molecule to the substrate-binding groove. There is a conserved and functionally important DFG motif at the N-terminal end of the activation loop.

#### 1.2.3.6 The P+1 loop

The P+1 loop is a small motif immediately downstream of the activation loop. Although its biochemical mechanism is not yet fully understood, it has been reported to play a role in recognizing the residues next to the tyrosine to be phosphorylated in the substrate [48,49]. This loop contains the conserved APE motif.

#### 1.2.3.7 The substrate-binding groove

A common feature of all protein kinases is that they phosphorylate polypeptide chains. The substrate-binding structure is the part of the kinase where the substrate proteins dock and it is formed by the catalytic loop, the P+1 loop and the $\alpha$D-, $\alpha$F-, $\alpha$G- and $\alpha$H-helices. This structure positions the phosphorylation site in the substrate protein next to the $\gamma$-phosphate of ATP.

#### 1.2.3.8 The ATP-binding pocket

The active site is located between the two lobes, the small N-terminal lobe above the ATP and the C-terminal lobe. The ATP-binding pocket has three differentiated parts: (1) a cluster of hydrophobic residues around the adenosine of ATP; (2) a set of charged residues around the $\gamma$-phosphate of ATP (namely the active site); and (3) a region of both polar and hydrophobic residues in the C-terminal lobe. From a mechanistic point of view, the hydrophobic area around the adenosine creates a binding pocket for ATP. The charged residues in the active site bind and guide the $\gamma$-phosphate and the divalent cation ($Mg^{2+}/Mn^{2+}$), and they participate in catalysis. The conserved residues in the C-terminal lobe help stabilize the whole pocket and probably, they also participate in the interaction with the substrate.

Five conserved residues are critical for ATP binding, active-conformation stabilization and catalysis (Figure 1.7). These are: (1) a lysine that interacts with the $\alpha$- and $\beta$-phosphates of ATP to stabilize it; (2) a glutamate in the $\alpha$C-helix that forms a salt bridge with this lysine, thereby increasing ATP stability; (3) an aspartate in the catalytic loop that is the catalytic base that seeds phosphotransfer by deprotonating the corresponding acceptor (Ser, Thr or Tyr) of the substrate peptide; (4) an asparagine that interacts with a secondary divalent cation, placing the $\gamma$-phosphate in the correct position; and (5) an aspartate from the DFG

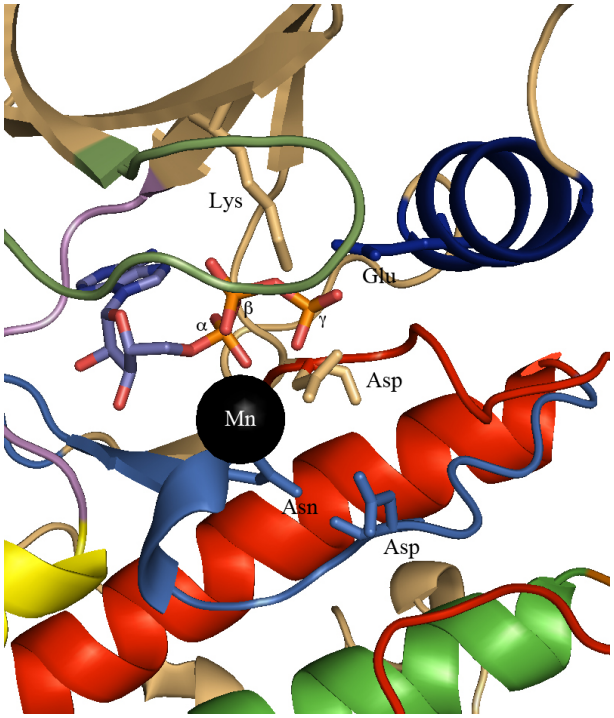



**Figure 1.7:** Detail of the ATP-binding pocket of human CDK2. The five highly conserved residues in the nucleotide-binding pocket [46] are represented by sticks. These are: (1) a lysine that interacts with the $\alpha$- and $\beta$-phosphates of ATP to stabilize it; (2) a glutamate in the $\alpha$C-helix that forms a salt bridge with this lysine, thereby increasing ATP stability; (3) an aspartate in the catalytic loop that is the catalytic base that seeds phosphotransfer by deprotonating the corresponding acceptor (Ser, Thr or Tyr) of the substrate peptide; (4) an asparagine that interacts with a secondary divalent cation, placing the $\gamma$-phosphate in the correct position; and (5) an aspartate from the DFG motif of the activation loop that chelates the primary divalent cation and hence, also positions the ATP molecule.

motif of the activation loop that chelates the primary divalent cation and hence, also positions the ATP molecule [46].

### 1.2.4 Allosteric control of the activity of protein kinases

As already mentioned, protein kinases play a key role in cells as they regulate important processes such as cell replication and signal transduction, among many others. Consequently, the activity and specificity of protein kinases must be tightly regulated without substantially interfering with the normal function of the proteins. Hence, protein kinases are often regulated by allosteric mechanisms that include effector binding, phosphorylation and protein domain or subunit interactions.

In the previous section, the structures commonly present in the PK domain were described. Allosteric regulation of protein kinase activity involves the coupling of these substructures through conformational rearrangements. Typically, this regulation involves spatial repositioning of the $\alpha$C-helix in the N-terminal lobe and the activation segment in the C-terminal lobe. The mechanisms that control these events differ greatly, despite the closely related sequences of the protein kinases and the similar active conformation adopted [50, 51].

The most common allosteric regulation is the change in the orientation of the DFG motif that is tightly coupled to the phosphorylation events occurring in the activation loop. In this conformation, the catalytic site is accessible to the nucleotide and the peptide substrates. Conversely, in the inactive phosphorylated conformation, the activation loop collapses into the active site and blocks substrate binding. Although slight differences exist, this mechanism is shared by a large number of proteins, including the insulin receptor kinase (IRK), c-SRC kinase, protein kinase A (PKA) and Abl kinase [51–56]. A phosphate usually interacts with a conserved arginine that plays a key role in the rotation of the DFG motif. Therefore, the whole catalytic loop is positioned in an optimal conformation for catalysis.

In addition to this activation mechanism, the $\alpha$C-helix regulates kinase inhibition in Src [54–56]. A Src homology 3 (SH3) domain interacts with a linker region between the protein kinase and the SH2 domains, resulting in the rotation of the $\alpha$C-helix and the disruption of the lysine-glutamate pair. Phosphorylation of a tyrosine in the C-terminal tail inhibits the activity of the protein by attaching the SH2 domain to the phosphorylated residue. This positions the SH3 domain such that it interacts with the linker region in between the other two domains, thereby inducing the linker to adopt a type II polyproline helical conformation. The SH3 domain and the linker can then stabilize the $\alpha$C-helix in the inactive conformation.

Another recurrent example in the literature is the control by regulatory subunit binding. For instance, CDK2 requires the binding of certain cyclins, mainly A and E, to the $\alpha$C-helix [51, 55]. The interaction with the cyclin forces the rotation of the helix and favors the reorientation of the DFG motif in the activation loop, and the interaction of the conserved lysine and glutamate important for the ATP-binding pocket [46]. The direct consequence of these rearrangements is the shifting of the activation loop, which in turn facilitates the interaction with cyclin-activating kinase (CAK). In Aurora kinase A, activation is boosted by binding of a microtubule-associated protein, TPX2, in addition to the phosphorylation of a threonine in the activation loop [51,57]. Similarly, Aurora kinase B binds to a centromeric protein, INCENP [58].

Dimerization is an alternative means of allosteric control [55], as exemplified by the epidermal growth factor receptor (EGFR) that must homodimerize with a second subunit to stabilize the active conformation [59]. Similarly, the interferon-induced, double-stranded RNA-activated protein kinase (PKR) must undergo back-to-back dimerization of the N-terminal lobes prior to activation [60, 61].

### 1.2.5 Accessory domains in protein kinases

As mentioned before, protein kinases rarely act alone but rather, they are often part of tightly regulated cascades. Signal transduction is generally modulated by a network of allosteric and phosphorylation events (including auto-phosphorylation), whereby a number of kinases activate their partners. While phosphorylation is mediated by the PK domain, other domains regulate protein kinase activity by associating them with other signaling modules or localizing the protein to the appropriate subcellular organelle. Studies [62, 63] in yeast have shown that kinases can be very promiscuous, phosphorylating a large number of different protein substrates while retaining remarkable specificity. This inconsistency suggests that kinases have a region committed to regulating the general function of catalysis, with another customizable region (or regions) that confer substrate specificity to the enzyme without any particular need to alter folding, compromise ligand binding or modify the subsequent reaction mechanism. This tightly regulated control mechanism is not unique to the protein kinase superfamily as many other protein families control their activity by interacting with accessory domains [64–66].

260 kinases contain no additional domains other than the PK domain, and many are indeed small proteins containing little more than this catalytic domain. For this reason, they are often controlled by additional regulatory (external) subunits, as exemplified by the cyclins when participating in controlling the cell cycle along with the activity of CDKs [23, 26].

By contrast, 258 out of the 518 human kinases in KinBase contain additional domains, and 83 different accessory domains other than the catalytic PK domain have been defined (Table 1.3). Although the general trend is that members of the same kinase family share the same domain structure, there are several cases where individual members of these families have gained (or lost) a domain, possibly yielding a slightly different function.

According to Manning *et al.* (2002) [40], the most common domains are those enabling kinases to interact with other signaling proteins: 25 proteins express SH2 domains that bind to phosphotyrosine residues; 38 have domains that are linked to small GTPase signaling proteins (RhoGEF, RhoGAP, RBD, PBD, RGS, CNH, HR1 or TBC domains); 42 have domains that are involved in lipid signaling (DAG_PE, C2, PX or PH domains); 28 carry domains that are associated with calcium signaling (CaM, IQ or OPR/PB1 domains); 7 proteins have domains that target proteins to the cytoskeleton (spectrin, cofilin, myosin head or FCH domains); and 49 kinases contain domains that mediate interactions with other proteins or RNA, such as the DEATH, SH3, SAM, LIM, ANK, RRM, DSRM or TUDOR domains.

| **Domain name** | **Genes** | **Doms.** | **Function class** |
|---|---|---|---|
| Protein kinase C terminal domain | 44 | 44 | Accessory domain |
| Immunoglobulin domain (Ig) | 30 | 254 | Extracellular, protein interactions |
| Fibronectin type III domain (FnIII) | 28 | 194 | Extracellular, protein interactions |
| SH2 domain | 25 | 27 | Adaptor, binds phosphotyrosine |
| SH3 domain | 27 | 28 | Adaptor, binds proline-rich motifs |
| PH domain | 23 | 22 | Signaling, phospholipid binding |
| Diacylglycerol binding (C1, DAG_PE) | 23 | 33 | Phospholipid binding |
| Calmodulin-binding motif | 23 | 25 | Not a Pfam domain |
| Sterile alpha motif (SAM) domain | 15 | 16 | Dimerization domain |
| Ephrin receptor ligand binding domain | 14 | 14 | Ligand binding |
| CNH domain | 12 | 12 | Putative cytoskeletal |
| Activin receptor | 11 | 11 | Ligand binding |
| HEAT, armadillo/$\beta$-catenin repeats | 10 | 27 | Protein interaction |
| Ankyrin repeat (ANK) | 9 | 59 | Protein interaction |
| p21-Rho-binding domain (PBD, CRIB) | 9 | 9 | GTPase interaction |
| Bromodomain | 8 | 13 | Acetyl-lysine (chromatin) binding |
| Regulator of G protein signaling (RGS) | 7 | 7 | GTPase interaction |
| PDZ/DHR/GLGF domain | 7 | 7 | Membrane targeting |
| Ubiquitin-associated domain A (UBA) | 7 | 8 | Protein degradation |
| Receptor L domain | 7 | 14 | Ligand binding |
| Furin-like cysteine rich region | 7 | 21 | Putative receptor dimerization |
| Phosphatidylinositol 3-kinase (PI3K) | 6 | 6 | Catalytic, protein kinase |
| FAT | 6 | 6 | Accessory domain for PI3K |
| FATC | 6 | 6 | Accessory domain for PI3K |
| Alpha kinase | 6 | 6 | Catalytic, atypical protein kinase |
| C2 domain | 6 | 6 | $Ca^{2+}$, phospholipid binding |
| Death domain | 6 | 6 | Dimerization domain |
| Guanylate cyclase catalytic domain | 5 | 5 | Catalytic, cGMP production |
| HSP90-like ATPase | 5 | 5 | Catalytic, atypical protein kinase |
| ABC1 family | 5 | 5 | Catalytic, atypical protein kinase |

**Table 1.3:** Most common Pfam domains in protein kinases apart from the PK domain. Adapted from Manning *et al.* (2002) [40]. Only domains present in at least 5 genes are shown.

## 1.3 Obtaining information about mutations

### 1.3.1 Resources that provide information about mutations

Currently, there is much effort being devoted to compiling and storing organized information about mutations. The annotations that usually accompany the mutations provide us with a powerful tool to understand the possible implications of the mutations in disease and, in the best case scenario, the biochemical mechanisms underlying the aberrant phenotypes.

Different databases provide a diversity of information that reflects their distinct goals, data sources, content, annotations and other issues. In fact, the scope of these databases is a key feature since some of the resources provide information about mutations from a general genome-wide perspective, while others provide detailed information on a protein family of interest. As an example of the latter, we will centre our attention on resources that annotate mutations affecting the protein kinase superfamily.

#### 1.3.1.1 dbSNP

The dbSNP database [67] (`www.ncbi.nlm.nih.gov/projects/SNP`) catalogues and annotates SNPs. Indeed, this resource constitutes the main source of information available on polymorphisms and their frequencies in different populations, sharing information with other state-of-the-art resources such as Ensembl [68]. There is no requirement or assumption about the minimum allele frequencies or functional neutrality for the polymorphisms in the database, although the current level of activity in terms of the discovery of general sequence variations suggests that SNP markers with unknown selective effects comprise most of the stored records. There is also no requirement about the species, although most of the mutations correspond to polymorphisms in humans or mice.

#### 1.3.1.2 Ensembl

As described in its latest release [68] (`www.ensembl.org`), Ensembl seeks to promote genomic science by providing high quality, integrated annotation on chordate and selected eukaryotic genomes within a consistent and accessible infrastructure. All the species included have comprehensive, evidence-based gene annotations and a selected set of genomes has additional data focusing on variation, comparative, evolutionary, functional and regulatory annotation. Currently, 56 species are supported including human and mouse.

#### 1.3.1.3 The HapMap project

The HapMap project [69] (`www.hapmap.org`) is a catalog of common genetic variants in the human genome. It describes these variants, their localization in the genome, and the allele frequencies among people within populations and among populations. The populations include individuals with African, Asian and European ancestry.

#### 1.3.1.4 OMIM Online Mendelian Inheritance in Man

OMIM [70] (`www.ncbi.nlm.nih.gov/omim`) is a curated collection of human genes and genetic disorders that was originally published as a book more than four decades ago. Afterward, this information is stored as free text. Information from the biomedical literature, including synopses and references is compiled manually. The lack of a structured database facilitates the flexible description of a wide range of phenotypes and genes, although the lack of standard format makes it more difficult to automatically search for information.

#### 1.3.1.5 SwissProt Variant pages and the ModSNP database

The SwissProt Variant pages [71] (`http://expasy.org/swissvar`) are an attempt to summarise the available sequence information in conjunction with the additional structural information on each human polymorphism in the SwissProt database [72]. The ModSNP database [71] maps mutations from the SwissProt Variant pages onto three-dimensional protein structures, using models generated by homology modeling when the referred proteins have not been crystallized and stored in the Protein Data Bank [73].

#### 1.3.1.6 SAAPdb

SAAPdb [74] (`www.bioinf.org.uk/saap/db`) is a resource for visualizing and analyzing the structural effects of mutations. The database classifies single amino acid polymorphisms into disease-associated mutations (*pathogenic deviations* : PDs) and *neutral polymorphisms* (SNPs), and it maps them onto structures in the Protein Data Bank [73]. In addition, the alteration of a set of structural features is assessed for each mutation to provide clues as to the pathogenicity of the variant allele. Neutral SNPs are mainly recovered from dbSNP [67] whereas pathogenic deviations are mainly gathered from OMIM [70].

#### 1.3.1.7 COSMIC: The Catalogue of Somatic Mutations in Cancer

The Catalogue of Somatic Mutations in Cancer [27] (`www.sanger.ac.uk/genetics/CGP/cosmic`) is a database of somatic mutations in four genes in which genetic alterations have been associated with cancer: BRAF, HRAS, KRAS2 and NRAS. The mutations, as well as tissue and histological data, are recovered manually from the literature and included after curation.

#### 1.3.1.8 KinMutBase

KinMutBase [75] (`http://bioinf.uta.fi/KinMutBase/`) is a plain-text database that stores disease-causing mutations within protein kinase domains. The first release of KinMutBase only contained information on protein tyrosine kinases while the second release also included, as well as an update of the tyrosine kinases, serine/threonine protein kinases. The current version contains 582 mutations in 20 tyrosine kinase domains and 13 serine/threonine kinase domains.

#### 1.3.1.9 MoKCa - Mutations of Kinases in Cancer

MoKCa [76] (`http://strubiol.icr.ac.uk/extra/mokca`) is a database that stores mutations in protein kinases that are likely to be involved in cancer. These somatic mutations in the protein kinase superfamily are recovered automatically from the literature, complemented with expert manual annotation and other predictions (such as the probability of being pathogenic), and then mapped onto the structures of the affected proteins.

### 1.3.2 Text mining techniques to extract kinase mutations from the literature

As discussed, several databases are devoted to the compilation, annotation and characterization of mutations occurring within the human kinome. Together they constitute a powerful resource to understand disease association and the functional/structural properties of the mutations that affect human protein kinases. However, compiling all the information available in databases and the literature remains a time consuming task, since there is no global repository covering all aspects of mutations.

One reason for this is that the sizeable amount of information provided by large-scale variation studies, and the growing efforts of databases and resources to store and curate this information, are still not perfectly and fully connected. The same applies for the many efforts dedicated to the detailed study of specific kinases in various biological systems published in individual research papers. Resolving this problem still generally involves the manual inspection and curation of specific variation studies, although this requires considerable resources.

Several text mining techniques can be applied to the automatic extraction of entities and their relationships from the existing literature, such as regular expressions, pattern recognition and natural language processing, among others. Indeed, these approaches have been successfully applied to other fields of research, for instance for the automatic extraction of protein-protein interactions [77, 78] and in the annotation of genes and proteins [79, 80]. Despite the success of these methods, it must be born in mind that this technology does not aim to replace manual curation and validation. Rather, text mining approaches are better understood as systematic tools to assist the efforts of human curators by helping them to find information, prioritize documents and highlight potentially relevant items [79, 81].

The consistent nomenclature used to describe mutations in the literature makes these entities especially amenable to this type of approach and accordingly, a growing number of such methods have been described in the literature over the years. A summary of several of these literature-mining tools to extract information on mutations is presented in Table 1.4.

| **Method** | **Main Features** | **Reference** |
|---|---|---|
| MEMA | Regular expressions, gene and protein mentions, co-mention proximity, OMIM validation. | Rebholz-Schuhmann *et al.*, 2004 [82] |
| MuteXt | Regular expressions, GPCR and NR mentions detection, co-mention proximity, sequence check. | Horn *et al.*, 2004 [83] |
| Yip | Regular expressions, protein mentions detection, Swiss-Prot validation, sequence check | Yip *et al.*, 2007 [84] |
| MutationGraB | Regular expressions, protein mentions detection, graph shorted distance, sequence check | Lee *et al.*, 2007 [85] |
| MutationMiner | Regular expressions, protein mentions detection, sentence co-mention | Baker and Witte, 2006 [86] |
| MuGeX | Regular expressions, protein mentions, protein and DNA mutation disambiguation | Erdogmus *et al.*, 2007 [87] |
| VTag | Machine learning detection of acquired sequence variation mentions detection (mutations, translocations and deletions) | McDonald *et al.*, 2004 [88] |
| OSIRIS | Detection of human gene variations corresponding to SNPs | Furlong *et al.*, 2008 [89] |
| MutationFinder | Regular expressions and patterns, protein mutations mentions detection, complex language expressions | Caporaso *et al.*, 2007 [90] |

**Table 1.4:** Summary of text mining implementations for mutation extraction.

### 1.3.3 Methods to predict the pathogenicity of mutations

We mentioned earlier that high-throughput resequencing screenings represent a powerful set of techniques to discover large numbers of mutations. Of these, only a small fraction are causally implicated in disease onset and therefore, separating the wheat from the chaff is still a major challenge [13]. For a small subset of the new mutations discovered, experimental information is available regarding the relationship between the mutation and disease, and for an even smaller number of cases the underlying biochemical mechanism is known. However, there is no information for the remaining mutations. The requirement of a lot of resources, time and money means that it is not feasible to experimentally test the association of all these mutations to disease, and to characterize their functional effects. Nevertheless, this problem is very amenable to *in silico* predictors.

Cline and Karchin [91] wisely described this approach as follows: *"A bench biologist interested in whether a mutation of interest impacts the transcription of a gene might perform site-directed mutagenesis on genomic DNA, transfect mutated DNA into cell culture and use readouts of the gene's transcriptional activity to measure changes with respect to wild type. In contrast, a bioinformatics approach typically involves computational analysis of the DNA sequence surrounding the mutation, possibly supplemented with information from published bench experiments."*

This is just one example of the very different methods available to predict the probability of a newly discovered mutation being implicated in disease. Several interesting reviews of this subject exist [13, 91, 92]. All predictors work under similar assumptions. Some make use of several features to highlight crucial positions in a given protein, and hence, rules are derived to predict the pathogenicity of mutations. Another group of methods assumes that evolutionarily conserved protein residues [93] are important for protein structure, folding and function, whereby mutations in these residues are considered deleterious. Variations on this principle lead to methods that predict deleterious mutations by evaluating changes in evolutionarily conserved Pfam motifs [94]. Moreover, a number of systems use protein structures to characterize substitutions that significantly destabilize the folded state. There are also methods that integrate prior knowledge in the form of both sequence-based and structure-based features from a set of mutations (previously characterized as pathogenic or neutral) to train an automatic machine learning system. After this training process, the system can infer the pathogenicity of new mutations based on the knowledge acquired. The most important predictors of pathogenicity are:

#### 1.3.3.1 SIFT

SIFT [93] is a method based on position-specific sequence conservation. The algorithm can be summarized as follows: (1) Search for similar sequences (PSI-BLAST); (2) Selection of closely related sequences that supposedly share function; (3) Multiple sequence alignment of PSI-BLAST results; (4) Calculate the normalized probabilities for all possible substitutions at each position from the alignment; and (5) Substitutions with normalized probabilities less than the cut-off (0.05) are predicted to be deleterious, otherwise they are tolerated.

#### 1.3.3.2 SNPs3D-stability

This method started as a rule-based system [95] but soon improved into a support vector machine (SVM) learning approach [96,97]. A number of features affecting binding and stability are analyzed:

- Protein stability.
  - Loss of hydrogen bonds: distance donor-acceptor less than 2.5Å or angle at the acceptor more than 90°.
  - Reduced hydrophobic interaction: loss of burial of more than 50Å$^2$ of non-polar area.
  - Loss of a salt bridge: two oppositely charged groups closer than 4.5Å apart.
  - Buried charge residue: introduction of an inaccessible charged residue.
  - Over-packing: introduction of a large residue chain.
  - Internal cavity: replacement of an inaccessible residue with a smaller residue chain.
  - Electrostatic repulsion: two equally charged groups less than 4.5Å apart.
  - Buried polar residue: polar groups with no accessibility and no hydrogen bonds.
  - Disruption of metal binding: replacement by non-metal binding.
  - Breakage of disulfide bonds: Cysteine replaced by non-cysteine.
  - Backbone strain: Glycine changed to another residue outside Ramachandran plot.
  - Backbone strain: Residue changed to proline with unfavorable torsion angles.
  - Backbone strain: cysProline ($\omega$=0±60°) changed to another residue.
  - Destabilization of a protein multimer: any of the above on a neighboring subunit.
- Ligand binding: any of the rules above where ligand atoms interact with the mutated side chain.
- Catalysis: the residue is involved in catalysis.
- Allosteric regulation: the residue is involved in allosteric regulation.
- Post-translational modification: disruption of N-X-S/T pattern for glycosylation.

#### 1.3.3.3 PolyPhen and PolyPhen II

PolyPhen [98] is a rule-based system that analyzes residue annotation, transmembrane and coiled coil regions, signal peptide propensities, incompatible replacement of amino acids (PSIC score), hydrogen bond disruption and hydrophobic or electrostatic interactions, disruption of ligand binding, disruption of secondary structure or exposure of the protein core, uncommon $\phi$-$\psi$ dihedral angles or untolerated regions of the Ramachandran map or normalized $\beta$-factors.

PolyPhen II [99] is an evolved version of the previous method, which relies on a *naïve* Bayesian classifier and a slightly modified set of features to predict the pathogenicity of mutations. The most informative of these characterize how the amino acids of interest are likely to occupy a given site according to the pattern of amino acid replacement in close homologues. Moreover, it assesses how distant the protein harboring the deviation is with respect to the protein from the human wild type allele, as well as whether the mutant allele originated at a hypermutable site. In total, 8 sequence-based features and 3 structure-based characteristics are evaluated.

#### 1.3.3.4 Panther

The Panther [100] algorithm uses evolutionarily related sequences to estimate the probability of a given amino acid at a particular position in a protein by calculating position specific evolutionary conservation (PSEC) scores. subPSEC is the difference of values resulting from the substitution of the wild type with mutant amino acids. The more negative the value, the more deleterious the mutation is likely to be, and subPSEC values smaller than -3 are considered deleterious.

#### 1.3.3.5 Pfam LogRE-value

The Pfam LogRE-value [94] score predicts whether a change will alter protein function by determining the difference in fit between the wild type version of the protein and a particular Pfam model. Positive LogRE values indicate variants that reduce the fit of the protein and that consequently, are potentially harmful.

#### 1.3.3.6 PMut

PMut [101] relies on a neural network to predict the pathogenicity of mutations. To train the system, the following features are assessed:

- Secondary structure.
- Solvent accessibility.
- Relative solvent accessibility.
- ProsaII differences in stability.
- Difference in surface potential.
- Difference in contact potential.
- Difference in the overall (weighted sum of the previous 2 potentials) potential.
- Blosum62 and PAM40 scores for the mutation.
- Residue volume change: van der Waals volume and volume of buried residues.
- Change in hydrophobicity: water/octanol free energy measurement and statistical potentials.
- Change in secondary structure propensities: Chou-Fasman and Swindells analyses.
- Sequence environment potential: difference in the probabilities of observing residue $r_j$ at position j given a sequence environment.
- Variability at the position: Shannon's entropy and average of scoring matrices.
- Changes in the associated position specific scoring matrix values.

#### 1.3.3.7 LS-SNP

LS-SNP [102] relies on an SVM implementing a radial basis function kernel. The following features are analyzed: differences in solvent, violation of spatial restraints, introduction of buried charges, evolutionary scores similar to subPSEC, Grantham values, changes in residue volume, hydrophobicity and formal charges.

#### 1.3.3.8 SNPs3D-profile

This sequence-based algorithm [103] uses a SVM to evaluate the probability of substituting the variant residue from the position specific scoring matrices (PSSM) , the conservation at that position of the alignment (Shannon's entropy), the mean entropy of the sequence, the standard deviation of the entropies and the entropy at each position as a Z-score.

#### 1.3.3.9 SNAP

SNAP [104] is probably the system that is currently producing the best results as demonstrated in the latest CAGI experiment (`http://genomeinterpretation.org`) It consists of a neutral network automatic machine learning predictor that evaluates the following features:

- Biochemical properties: Several properties are evaluated here. These include, the assessment of the introduction of charged residues into a buried positions, the presence of inflexible prolines in an $\alpha$-helix, the replacement of hydrophilic amino acids with hydrophobic amino acids (or *vice versa*), the assessment of over-packing or the presence cavities in the protein core, changes in $C_\beta$-branching and the calculation of the mass of the wild-type and mutant residues.
- Amino acid type.
- Transition frequencies.
- PSI-BLAST profiles (similar to subPSEC).
- PSIC score: Absolute PSIC score of wild-type and mutant residues, difference (subPSIC) between wild-type and mutant PSIC score and a 3 state PSIC score (cut-offs at 0.5 and 1.5) are evaluated at this step.
- Secondary structure (PROFsec).
- Relative solvent accessibility (PROFacc).
- Predicted flexibility (PROFbval).
- PFAM annotation: Different properties are considered. The presence of domain boundaries at the position, the model score of the domain, whether the position is conserved or whether the mutation is a better match with a consensus domain, the presence of domains in the area are included.
- SwissProt annotation: Active residues, bonding residues, post-translational modifications, variable residues and transmembrane regions are considered in the calculation.
- SIFT.
- PolyPhen.

#### 1.3.3.10 CanPredict

CanPredict [105, 106] is a random forest classifier that analyzes the scores from SIFT, Pfam-based Log RE-value and Gene Ontology Similarity Score (GOSS) to provide a meta-prediction of (cancer) pathogenicity.

1. SIFT: As described previously, this method uses the similarity between closely related proteins to identify potentially deleterious changes (deleterious if SIFT score smaller than 0.05).

2. Pfam-based logRE-value: As described previously, score predicts whether a change will alter protein function by determining the difference in fit between a wild-type version of the protein and a particular Pfam model.

3. GOSS: uses Gene Ontology (GO) log odds to measure the similarity of the submitted protein to other known cancer-causing genes. The sum of individual scores for each GO term is considered. And each individual GO score is calculated as the logarithm of the percentage of disease genes annotated with that GO respect to the percentage of neutral genes annotated with that GO term.

#### 1.3.3.11 Torkamani's kinase specific predictor

This method [107] corresponds to an SVM implementing a radial basis function kernel. The main particularity of this predictor is that it is trained specifically for the protein kinase superfamily. Indeed, the training vectors include features that are specific to this superfamily, such as membership to a particular kinase group, which gives it a high discriminative power. The complete list of features evaluated includes:

- Membership to a given kinase group.
- Wild type and mutant amino acids.
- Membership of the kinase domain, N-terminal and C-terminal lobe.
- subPSEC score (as in Panther).
- Change in biochemical properties: hydrophobicity, polarity and charge.
- Secondary structure prediction (Proteus server).
- Solvent accessibility (accessible, inaccessible and intermediate using Predict Protein).
- Change in flexibility.
- Difference in Kyte-Doolittle hydropathy.
- Difference in water/octanol partition energy.
- Difference in volume.

#### 1.3.3.12 SNPs&GO

SNPs&GO [108] is a genome-wide predictor of the pathogenicity of mutations. As most of the best performing classifiers, it implements a SVM that relies on a radial basis kernel function for the classification. The features implemented are completely sequence-based, which improves the scope of the predictor, particularly for those proteins for which a crystal structure is not yet resolved. The features analyzed include:

- Wild type and mutated amino acids
- Sequence environment (20 slots, accounting for the occurrence of each amino acid type in a 19-residue window centered on the mutation at hand (9-x-9).
- Profile information: analysis of the conservation of the position in a multiple sequence alignment. This is calculated as the frequency of the wild-type and mutant residues, depth (number of sequences) at this position and conservation index.
- PANTHER prediction (as described before): Properties include the disease probability of the mutation from Panther, probability of find the wild-type and the mutant residues and the number of independent counts (a measure of the global diversity of the sequences at this position).
- Gene Ontology Log-Odds ratio.

#### 1.3.3.13 MutationAssessor

MutationAssessor [109] is a genome-wide predictor of the functional impact of non-synonymous mutations. The prediction is based on earlier calculations of the differential evolutionary conservation of amino acids within protein subfamilies [110]. Protein subfamilies are determined by clustering multiple sequence alignments of homologous sequences on a background of overall function conservation.

#### 1.3.3.14 Condel

Condel [111] is a genome-wide meta-predictor of the functional impact of missense mutations. As a meta-predictor it integrates the output of 5 different methods: SIFT [93], MutationAssessor [109], PolyPhen2 [99], Pfam LogRE-value [94] and MAPP [112].

Chapter 2

# Goals and Objectives

## 2.1 Motivation

This doctoral thesis will focus on the analysis of mutations in the protein kinase superfamily. Protein kinases are a good subject of study for a number of reasons. First, because of the completeness of the information about their biochemical / cellular mechanisms. Second, kinases are a large and complex superfamily, hence, they provides a rich source of information for comparative analysis. And, third it is an important family, many of its members play a crucial role in cancer and other diseases.

Most of the mutations published in the literature are tolerated and consequently, they are apparently neutral to the cell. By contrast, only a small number of them are causally associated with human diseases and cancer. The mechanisms by which pathogenic kinase mutations lead to aberrant phenotypes have been explored in a number of relevant cases for which the mechanism of action are relatively well understood. However, the biochemical characterization of each single mutation requires an enormous effort, involving the dedication of time and resources, which cannot keep up with the pace of high-throughput mutation discovery technologies.

Thus, there is a clear need to broaden our understanding of the mechanisms by which some mutations alter protein kinase function and cause disease. Such advances will help to develop cost-efficient characterization protocols and, in parallel, prioritize the study of mutations focusing on those that are most likely to play a direct role in human disease.

I have approached these problems using Computational Biology tools and methods. CB provides a powerful framework to analyze vast amounts of complex and heterogeneous data. Furthermore, CB methodologies are used to generalize previous observations and to use them to predict the potential pathogenicity of newly discovered mutations.

In summary, the ultimate goals of this thesis are to understand the mechanisms by which pathogenic mutations disrupt the structure and function of expressed protein kinases, and to design a reliable methodology to identify mutations causally implicated in human disease.

## 2.2 Specific Objectives

This thesis is subdivided into four independent sections:

1. *Mutations mapped onto the tertiary structures of proteins.*
   Here, we will tackle the technical problems of the integration of information from diverse sources into a common repository and subsequently, transferring this primary sequence information to the tertiary structures of proteins. A public web-server, 3DSim, will display the mutations and their annotations mapped onto representative structures. A public web-server, 3DSim, will display the mutations and their annotations mapped onto representative structures.

2. *Text mining techniques to enhance knowledge of the kinase mutations.*
   The objective here is to compile information from existing data sources and in particular, to automatically extract mutations from the literature. In addition, we will associate the mutations with their original manuscripts, providing a summary of the evidence in the literature that will help to link mutations with disease phenotypes.

3. *Characterization of the residues that disrupt protein kinase function.*
   The goal is to identify the properties that distinguish pathogenic and neutral mutations. For that, we will compare the distribution of selected features between neutral and potentially pathogenic mutations, characterizing those that could be relevant to pathogenicity.

4. *Prioritization of disease-associated kinase mutations.*
   The objective of this final task is to develop an automatic system to predict the pathogenic nature of mutations based on the distribution of selected features. The new method will be benchmarked against previously published approaches and it will be made publicly available as a web server to facilitate the analysis of newly detected mutations.

Chapter 3

# Materials and Methods

## 3.1 Mapping mutations from the sequence to the proteins' tertiary structure

### 3.1.1 Obtaining mutations from SAAPdb

SAAPdb [74] is a database of single amino acid polymorphisms (SAAPs) from several resources, such as dbSNP [67], ADABase [113], G6PD [114], HAMSTeRs [115], p53 Database [116], LDLR [117], OMIM [70], OTC [118], SOD1db [119] and ZAP70Base [113]. Where possible, these polymorphisms have been mapped to protein structures in the PDB [73]. As of October 2008, SAAPdb contains 9,060 unique pathogenic deviations (PDs: SAAPs associated to a disease) and 2,532 unique single nucleotide polymorphisms (SNPs: SAAPs with no known pathogenic effect) successfully mapped to the UniProtKB sequences in Gene3D [120].

Pathogenic deviations and neutral SNPs were only taken into account if the alteration introduced is not silent, i.e. if it produces a change of amino acid from the wild-type. Since SAAPdb integrates annotations from different sources there may be discrepancies in the annotations for some mutations. These inaccuracies in the annotation are incompatible with the analyses performed here and consequently, the mutations that were annotated as both neutral and disease-associated were all considered as pathogenic deviations.

### 3.1.2 Generating groups of Gene3D sequences represented by the same CATH domain

All CATH domains were recovered directly from the database for each of the 2,097 superfamilies available in CATH [121] at the time of analysis (release 3.2 August, 2008), along with the corresponding amino acid sequences. The sequences were grouped together by superfamily and a total of 86,463 CATH domains were retrieved.

In order to assign the closest CATH domain to each of the Gene3D sequences in a given CATH superfamily, we queried each of the sequences against the database of CATH domains generated previously for that superfamily. Queries were performed using BLAST [122] and the best match for each of the Gene3D sequences was considered the closest CATH domain. Only cases where the identity between the hit and the query was greater than 20% were considered and the resulting CATH domain was assigned as the structural representative of this sequence. After performing this classification for the whole set of sequences, all Gene3D

sequences represented by the same CATH domain were grouped together and all the sequences within a group were considered to be similar. A total of 2,091 unique groups were generated.

### 3.1.3 Mapping SAAPdb mutations to representative CATH domain structures

During the previous step of the pipeline, the sequences represented by the same CATH domain were considered similar. However, in order to collapse all the mutations from the Gene3D sequences onto the representative CATH domain sequences, the equivalence between pairs of residues should be established. To perform this task the sequences were aligned using the MUSCLE [123] package and the resulting alignments were used to transfer the mutations, both pathogenic deviations and neutral SNPs, from the sequences in Gene3D to the corresponding representative CATH structures.

## 3.2 Characterization of the residues that disrupt protein kinase function

### 3.2.1 Classification of the human kinome according to KinBase

The KinBase [40] resource (`www.kinase.com/kinbase`) is a repository storing the currently accepted classification of eukaryotic protein kinases. At the time of the analysis, KinBase contained 620 human protein sequences, excluding pseudogenes, of which 518 correspond to protein kinases. KinBase does not directly map its entries to Uniprot and consequently, we used BLAST to search each KinBase sequence against a custom database containing all human sequences annotated with a protein kinase domain in Uniprot [72]. After this mapping step, we were able to assign 488 KinBase identifiers to a valid Uniprot entry, 474 of them (97.13%) at sequence identity levels $\geq$95%.

### 3.2.2 Selection and classification of somatic mutations

Large-scale systematic genotyping studies have been carried out to identify mutated genes in human cancer genomes. These studies focused on different tumor types (Colon and breast) or on the protein kinase superfamily [14, 16], and they led to the identification of more than 3,500 somatic mutations. Further statistical analysis based on mutation rates or synonymous versus non-synonymous mutation ratios have been used to classify genes as *drivers*, those causally involved in oncogenesis, or *passengers*, those incidentally involved in oncogenesis and that probably arise during tumor development.

The analysis performed in this doctoral thesis focuses on the protein kinase catalytic domain. The protein kinase domain subset comprises 140 mutations, 73 (52%) drivers and 67 (48%) passenger mutations. We stored these mutations along with the corresponding amino acid change, their nature as a driver or passenger, and their sequence and structural positions. The mapping of sequence and structure positions was calculated with the SPICE DAS server [124], a resource originally developed to visualize sequence-based annotations in protein structures.

### 3.2.3 Selection and classification of germline mutations

SAAPdb [74] is a database of single amino acid polymorphisms (SAAPs) mapped to protein structures. SAAPdb aims to define potential structural consequences of mutations and to identify differences in the structures that may be associated with neutral and pathogenic mutations.

The disease dataset, pathogenic deviations, is mostly derived from OMIM [70], whereas the neutral dataset comes from dbSNP [67]. Within the Protein Kinase domain of the 488 kinases we could map to Uniprot identifiers, SAAPdb contains 130 pathogenic deviations and 200 neutral SNPs. Of these 130 PDs, 62 were successfully mapped to a residue in a resolved PDB structure. Similarly, of the 200 SNPs mapped to the sequence, 36 were mapped to a PDB structure. A unique mutation was defined by the combination of four parameters: UniProtKB accession number; sequence position; native amino acid; and mutated amino acid. We excluded nonsense and synonymous mutations assuming that they have a known truncating effect or no effect on the protein structure, respectively.

### 3.2.4 Calculation of sequence conservation

Two different measures of sequence conservation were used.

First, for each position in the multiple sequence alignments, conservation was measured in terms of Shannon's entropy [125], a measure of the variability in the distribution of elements in a set described by the formula:

$$-\sum_{i=1}^{n} p(x_i) \cdot log_2 p(x_i)$$

Where $p(x_i)$ is the probability of having element $x_i$ in bin $i$ for that distribution.

Conservation was measured in the context of identity, using 1 bin for each amino acid and an extra bin for gaps, giving a total of 21 bins. The positions in the alignment were labeled as conserved if the Shannon's entropy was $<0.20$. Positions with more than 75% gaps in the corresponding multiple sequence alignments were labeled directly as unconserved.

Second, we used the weighted sequence conservation implemented in AL2CO [126]. AL2CO is a program that calculates a conservation index for each position in a multiple sequence alignment using diverse methods that share a common principle, whereby amino acid frequencies at each position are estimated and the conservation index is calculated for these frequencies. We used the implementation to weight sequences in order to correct for the unequal distances between different sequence pairs in the alignment and the matrix score that gives the most common position occupied by residues with similar physicochemical properties. Those residues with a normalized conservation index of $\geq$70% were considered to be conserved.

### 3.2.5 Calculation of the solvent accessibility with Naccess

Naccess (Hubbard *et al., unpublished*) is a stand-alone program that calculates the solvent accessible area by rolling a probe with a van der Waals radius over the surface of the molecule. We defined residues as buried if their relative accessible surface area exposed to the probe is less than or equal to 16% of the total surface of the residue.

### 3.2.6 Defining the ATP binding site with FireDB

The FireDB database [127] contains a comprehensive curated set of substrate-binding and catalytic residues, extracted directly from PDB [73] or from the Catalytic Site Atlas [128]. FireDB binding residues for the various kinases were mapped into the general model using the corresponding multiple structure alignment.

### 3.2.7 Defining specificity determining positions with S3Det

Specificity Determining Positions (SDPs) are the positions within a family of homologous proteins characteristic of the internal organization into subfamilies. Hence, SDPs typically correspond to positions in a multiple sequence alignment where the type of amino acid is differentially conserved among subfamilies. Subfamilies and SDPs have previously been associated with important features of protein functional subspecificity, such as catalytic activity, small ligand binding, protein-protein interactions, DNA/RNA binding, etc. [129–131]

Here we predict SDPs with S3det [131], an in-house novel implementation of the Sequence Space approach [129] that relies on Multiple Correspondence Analysis [132]. The procedure implemented in S3det is summarized in Figure 3.1.

Using this technology, S3det vectorizes the initial multiple sequence alignment in order to produce two linked spaces, namely a protein space and a residue space. These two spaces can be represented simultaneously: in the protein space similar sequences are located in the same region of the space, whereas the residues characteristic of those sequences map to an equivalent region in the residue space. Through automatic clustering in the sequence space, S3det detects groups of sequences that correspond to the family's subfamily composition. These subfamilies are then mapped to the residue space through the vector that represents the center of their masses. This procedure permits the set of residues that uniquely characterize each subfamily to be determined automatically, defining the SDPs. SDPs determine the segregation into subfamilies and they are responsible for the specific functional features of the subfamily.

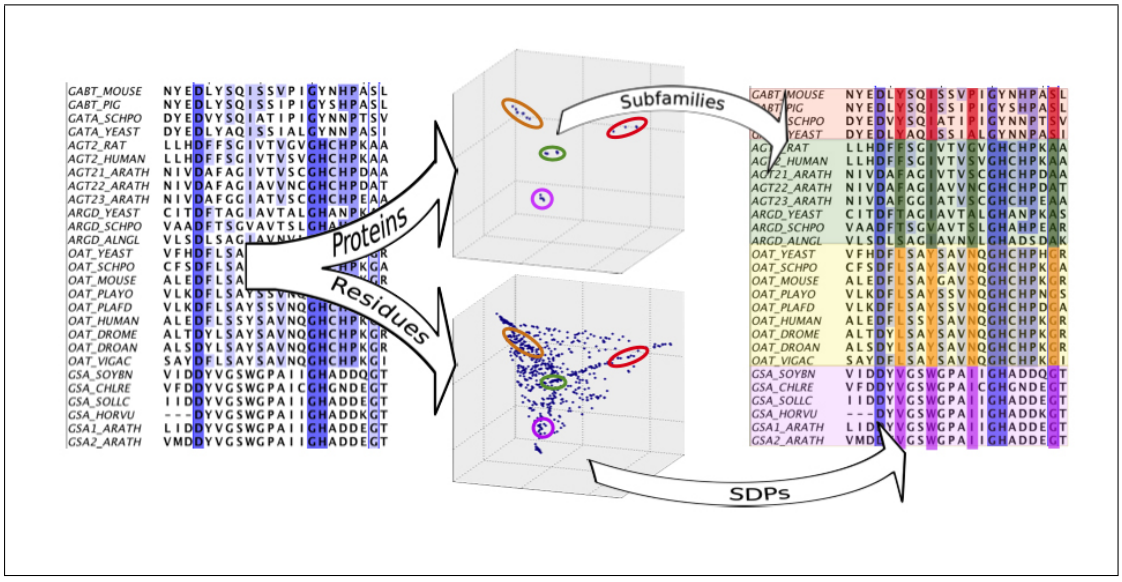



**Figure 3.1:** Summary of the method to identify specificity determining positions (SDPs) implemented in S3det. [131]

### 3.2.8 $Xd$ analysis

To assess the significance of the proximity of different sets of mutations to specific areas of the protein (buried, functional, conserved, etc), we used the harmonic deviation introduced previously, $Xd$ [133–135]. The most relevant characteristic of this measure of the differences between distributions is that it more heavily weights those positions in bins proximal to the

features studied, instead of considering the complete distribution of distances to be equally informative. In other words, it gives more importance to differences in the distribution of residues close to the important regions than to differences located in more distant regions.

$$Xd = \sum_{i=1}^{n} \frac{P_{ic} - P_{ia}}{d_i \cdot n}$$

Where $n$ is the number of distance bins in the distributions, $d_i$ is the upper limit for each bin, $P_{ic}$ is the percentage of residues with a distance between $d_i$ and $d_{i+1}$, and $P_{ia}$ is the same percentage for all the residues in the protein.

Defined in this way, positive values of $Xd$ would indicate that the population of residues shifts to smaller distances with respect to the population of all residues. In practice, we use an arbitrary value of 0.5 for the difference of $Xd$ ($\Delta$Xd) to indicate distributions of residues that are sufficiently distinct in terms of their proximity to the chosen areas of the protein.

## 3.3 A prediction server to determine the pathogenicity of a mutation

### 3.3.1 Mutation Dataset

The mutation data used here was derived directly from Uniprot [72] (release 2011_01; Jan 11, 2011) after applying the following constraints:

1. The protein is annotated as a protein kinase in UniProt.
2. It is a human protein.
3. The mutation corresponds to non-synonymous, non-truncating single point coding mutations. Other mutation types such as insertions, deletions, copy number alterations, truncating and silent mutations were not considered in this analysis.

The use of a Uniprot derived dataset has recently been benchmarked for a number of classifiers with satisfactory results [136].

Following this pre-filtering step, we classified the mutations as disease or neutral mutations according to the annotation in Uniprot. There is a third group in Uniprot that aggregates the mutations for which insufficient information is available, mutations that were ruled out of this analysis. After the whole selection process, the 'disease dataset' that includes mutations for which there is experimental evidence of their disease association, contained 865 mutations in 65 human kinases. By contrast, the 'neutral dataset' that contains mutations with no experimental proof of association to disease, contained 2,627 mutations in 447 human protein kinases.

### 3.3.2 Implementation of the classifier

To implement the Support Vector Machine classifier we used the SVMLight (`http://www.cs.cornell.edu/people/tj/svm_light/`) package with a radial basic function (RBF) kernel:

$$K(x_i, x_j) = exp(-G||x_i - x_j||^2))$$

In this manner, two parameters are crucial to the performance of the classifier: the soft-margin penalty (C) and the radius ($\gamma$). An exhaustive evaluation of the parameters was carried out for

values ranging between $0 \leq C \leq 8$ in 1 unit steps, and $10^{-4} \leq \gamma \leq 10^{-2}$ increasing by $5 \cdot 10^{-4}$ after each run. In order to decide which pair of parameters performs best, the average f-score across the entire set of k-folds was chosen as the scoring function for the optimization.

### 3.3.3 Evaluation of performance

The classification obtained from our classifier was evaluated using a 10-fold cross-validation approach where 80% of the mutation data is used to train the classifier, another 10% to validate and optimize the parameters, and the remaining 10% is used to evaluate the classifier. The process is repeated enough times to ensure that all subsets of mutations are used for each purpose. The classifier's performance is averaged across all combinations in order to avoid over-interpreting the quality of the method. The efficiency of the classifier can be assessed in many ways and here we describe the most illustrative ones.

Hereafter, we will refer to the following abbreviations:

**TP** True positives, correctly predicted disease-associated mutations.

**FP** False positives, neutral mutations predicted as disease prone.

**TN** True negatives, correctly predicted neutral mutations.

**FN** False negatives, disease-associated mutations predicted as neutral.

Accuracy, often referred to as $Q_2$, accounts for the fraction of mutations correctly predicted in function of the total number of mutations.

$$Accuracy = \frac{Correctly\ Predicted}{All\ Predicted} = \frac{(TP + FN)}{(TP + TN + FP + FN)}$$

Recall, also referred to as sensitivity by other authors, accounts for the proportion of correctly predicted disease-associated mutations in function of all the disease-associated mutations in the dataset.

$$Recall = \frac{Correctly\ Predicted\ (disease)}{Observed\ Mutations\ (disease)} = \frac{(TP)}{(TP + FN)}$$

Precision accounts for the proportion of correctly predicted disease-associated mutations with respect to all the predicted disease-associated mutations.

$$Precision = \frac{Correctly\ Predicted\ (disease)}{Predicted\ Mutations\ (disease)} = \frac{(TP)}{(TP + FP)}$$

The F-score is a measure of the accuracy of the classification. It considers both the precision and the recall in a single representative score for evaluation purposes.

$$F - score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

The Matthews Correlation Coefficient (MCC) was calculated according to the following formula:

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

### 3.3.4 Classification Feature: membership to a Kinase group

In order to cluster the kinases according to the groups they belong to, two different classification schemes were used. The KinBase resource [40] constitutes the currently accepted classification scheme of eukaryotic protein kinases. According to KinBase, kinases are categorized as 'conventional' protein kinases (ePKs) or 'atypical' protein kinases (aPKs). The ePKs form the largest group and they have been subdivided into 8 groups according to sequence similarity, the presence of accessory domains and by considering different modes of regulation. The eight ePK groups defined in KinBase correspond to: the AGC group (including cyclic-nucleotide and calcium-phospholipid-dependent kinases, ribosomal S6-phosphorylating kinases, G protein-coupled kinases and close relatives of these kinases); the CAMKs (calmodulin-regulated kinases); the CK1 group (casein kinase 1 and close relatives); the CMGC group (including cyclin-dependent kinases, mitogen-activated protein kinases, CDK-like kinases and glycogen synthase kinase); the RGC group (receptor guanylate cyclase kinases); the STE group (MAPK cascade kinases); the TK (tyrosine kinase) and the TKL (TK-like), which are a group of serine-threonine kinases resembling TKs. Another broad miscellaneous group, called 'Other', is also considered for those proteins that do not fit in any of these predefined categories. By contrast, Uniprot [72] provides a classification scheme that includes the same groups included in KinBase along with the additional groups, NEK and STG, making a total of 11 groups. The vector of features submitted to the classifier contains a position for each of the groups in the latter scheme. The values are encoded as 1 for the group to which the kinase housing the mutation belongs to, and 0 for the rest. A similar approach was followed by Torkamani and Schork [107].

### 3.3.5 Classification Feature: Gene Ontology Log Odds Ratio

The Gene Ontology Log Odds Ratio (GOLOR) was used to classify the mutations as pathogenic (disease-associated) or neutral according to the annotations regarding the function of the genes in which they exist. To compute the score, we retrieved all the terms associated to the kinases in our dataset from the 3 sub-ontologies in Gene Ontology [137] (Molecular Function, Biological Process, Cell Component). The ontologies were followed towards the root of each ontology in order to include all parental terms in the calculation. Note that 'part-of' relationships were discarded and only 'is-a' links were considered. For each of the kinase genes, the sum of the Gene Ontology Log Odds Ratio (sumGOLOR) was computed as follows:

$$sumGOLOR = \sum log_2 \frac{\%\ kinase\ genes\ annotated\ with\ GO_i\ in\ the\ disease\ set}{\%\ kinase\ genes\ annotated\ with\ GO_i\ in\ the\ neutral\ set}$$

Where disease-associated kinase genes are those with at least one reported disease-associated mutation and neutral kinase genes are those with no reported disease-associated mutation. In order to resolve undetermined ratios, frequencies equal to 0 were artificially set to $10^{-9}$. A similar approach with slight changes in the algorithm is followed in two other methods: CanPredict [105, 106] and SNPs&GO [108].

### 3.3.6 Classification Feature: PFAM domains

The position of the different domains in the sequence of the human protein kinome was extracted from the swisspfam file in PFAM [138]. A binary position in the vector was created for each of the 117 different domains in the protein kinase family, where 1 means that the mutation is in a position that was characterized as part of that domain, and otherwise it is attributed a value 0. An additional binary position in the vector, pfam_any, was created to record whether the position belongs to at least one PFAM domain. This is a simplified version of the implementation by other authors [104, 105, 107].

### 3.3.7 Classification Feature: Amino acid type and change in hydrophobicity

Each amino acid type was encoded at 20 positions in the vector, where the wild-type residue is encoded as 1 and its mutant counterpart is encoded as -1. The rest of values remain as 0 for classification purposes. An additional position was encoded to represent the change in the Kyte-Doolittle hydrophobicity index [139].

### 3.3.8 Classification Feature: Uniprot Annotation

Uniprot [72] provides a detailed description of the residues for a number of proteins in the database. We considered 5 different classes of residue annotation to be relevant:

1. Catalytic site (including residues annotated as SITE, BINDING, ACT_SITE, METAL and NP_BIND: refer to the Uniprot help pages for a detailed description of the annotations)
2. Disulfide bond (DISULFID)
3. Post-Translational Modifications (MOD_RES, SIGNAL)
4. Residues with special interest (MUTAGEN)
5. Transmembrane regions (TRANSMEM).

A binary input corresponding to each of these categories was added to the classification vector. In addition, two additional positions were added: one that corresponds to a positive match in at least one category from the catalytic site class, while the other corresponds to a positive match in at least one of the categories described above. A similar approach was followed previously [98, 104, 140, 141].

### 3.3.9 Classification Feature: Phosphorylation sites

PhosphoELM [142] is a database of eukaryotic phosphorylation sites. This resource includes manually curated information derived from the literature, as well as high-throughput analyses for 1,232 phosphorylation sites in 287 human kinases present in our dataset. Since out of the 20 potential residues only 3 can be phosphorylated (Ser, Tyr, Thr), this feature was encoded for the classifier according to 3 different states: 1 represents a reported residue amenable for phosphorylation, 0 if the residue is a Ser, Thr or Tyr that is not phosphorylated, and -1 for the remaining residues.

### 3.3.10 Classification Feature: Catalytic sites

FireDB [127] is a database of known functionally relevant residues. It includes both biologically relevant data filtered from the close atomic contacts in 3D crystal structures and manually annotated catalytic residues. The presence of a mutation in the catalytic site of a protein was encoded in the classifier as a binary input, whereby 1 means that the mutation is part of the catalytic site, 0 otherwise.

### 3.3.11 Classification Feature: Evolutionary Information

In order to capture the similarity between closely related proteins and thereby identify potentially deleterious changes, we included the SIFT score in our feature vector [93]. This method relies on the normalized probabilities for all possible residue substitutions at each position of a multiple sequence alignment of homologous proteins. The score can be easily translated into a binary output where values $<0.05$ are considered deleterious. Consequently, both the binary

and the continuous versions of the score were computed in order to provide more discriminating results. Additionally, the number of sequences in the alignment at the position of interest was also considered. Since it was introduced in 2001, this method has been successfully incorporated in several predictors of pathogenicity [104, 105, 107, 141].

### 3.3.12 Classification Feature: Specificity Determining Positions

Those positions occupied by conserved residues within groups of proteins in a family sharing a common general specificity that differs between groups can be used as a proxy for the regions accounting for subfamily specificity. SDPs, also referred as tree-determinants on occasion, were calculated using a simplified version of the in-house S3Det predictor [131]. In our implementation, the f-score associated to the wild-type and mutant residues in the classification of the subfamilies calculated from the sequences in the PFAM alignments was encoded in the classification vector. An additional third position represented the difference between these two scores. This difference represents the change (increase or decrease) in agreement with the subfamily introduced by the mutation.

Chapter 4

# Results

## 4.1 Mapping mutations: From the protein sequence to its tertiary structure

This thesis was conceived in order to address the need to transfer information about mutations from databases and other sources to the corresponding protein sequences and structures. We needed a system that could handle the heterogeneous information available regarding mutations and, very importantly, a system that could transfer the coordinates of the mutations and the associated meta-information between protein sequences, and their corresponding structures and structural families.

As a framework for our development we used the CATH database developed by the group of Christine Orengo at the University College London [121]. CATH is a manually-curated hierarchical classification of protein domain structures. In CATH each protein structure is partitioned into domains and assigned to superfamilies, with each superfamily representing groups of domains that share a common evolutionary origin. CATH is coupled to Gene3D [120], a resource that assigns protein sequences to the CATH hierarchy and that fills the sequence-structure gap. In the CATH framework, the representative structure of each CATH domain is a proxy of all the associated sequences. Therefore, the system makes it possible to fulfill our need, transfering the coordinates of the mutations and the associated meta-information between sequences and structures of the same superfamily.

### 4.1.1 3DSim in a nutshell

3DSim [143] is a method to map single amino acid polymorphisms onto protein structures. In the current implementation (See Introduction for a detailed description) mutation data is obtained from SAAPdb [74], sequence information from Gene3D [120] and the protein structures correspond to the superfamily representatives in CATH [121]. The system permits the mutations to be mapped onto the protein structures and it facilitates the dynamic visualization of the mutations within the structures. Relevant annotations about the mutations are summarized, including information regarding the wild type and mutant sequences, the predicted structural implications of the mutations and their characterization as neutral or pathogenic according to these predictions. The information is provided together with the links to the original sources of information, such as UniProt, Gene3D, SAAPdb, CATH, Modbase, etc...

The algorithm behind the server is described in detail in Figure 4.1 and it can be summarized as follows:

1. Each mutation in SAAPdb (either neutral or pathogenic) was located in the corresponding Gene3D sequence, which acts as a mediator between the sequence and structure databases. The mapping between SAAPdb and Gene3D included 11,904 sequences.

2. Gene3D sequences were clustered according to their closest representative CATH domain based on homology. This ensured that all sequences were associated with a unique CATH domain, even if their own protein structures were not resolved. The closest relative found was used to cluster the sequences (i.e., the one with the lowest e-value and highest sequence identity in a BLAST search). This process yielded 2,091 different groups.

3. All sequences corresponding to the same CATH domain were aligned to the sequence of the CATH domain's representative using the multiple sequence alignment package MUSCLE [123].

4. The multiple sequence alignments were used to identify the equivalent sequence to structure positions, which were then used to transfer the mutations in SAAPdb from each of the Gene3D sequences onto the structures of corresponding CATH domain representatives.

At the end of the whole pipeline, we collected information on 6,514 point mutations; 4,865 of them are known to be associated with disease and map to 396 CATH superfamilies.

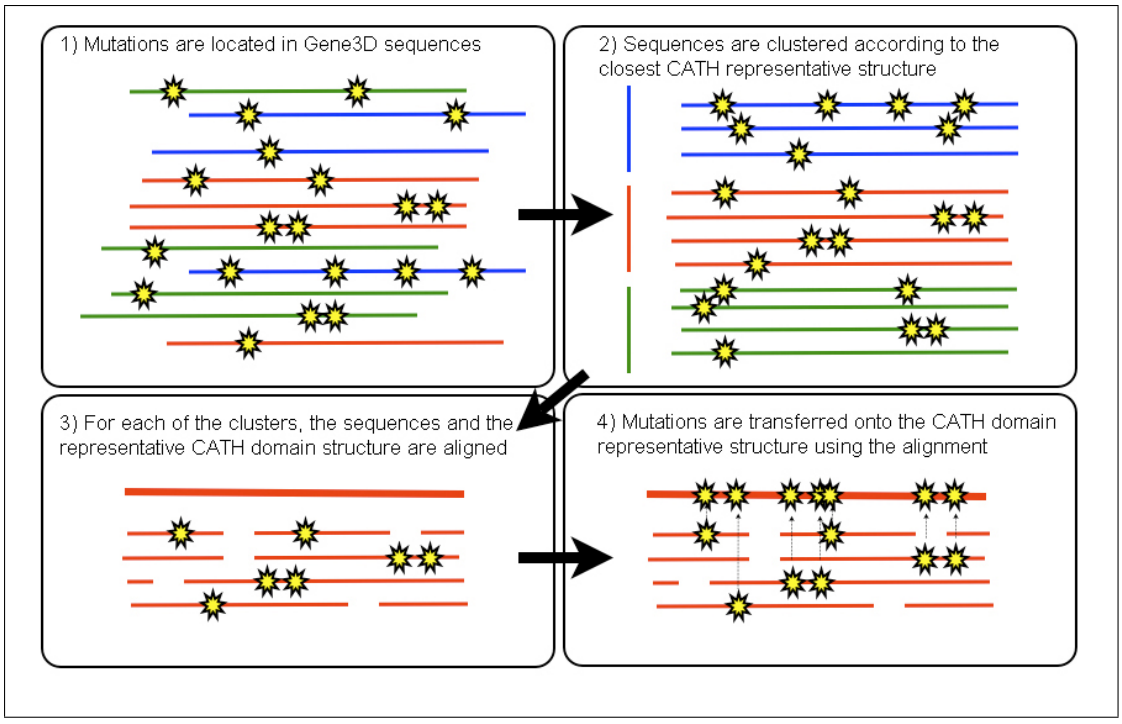



**Figure 4.1:** Schematic view of the algorithm that 3DSim uses to map mutations in SAAPdb from the sequences onto the structures of the family representatives in CATH.

### 4.1.2 Implementation of the application as a web server

To make the information generated available to the scientific community, we decided to implement our mutation sequence-to-structure mapping pipeline as a web server. Thus, 3DSim [143] is publicly available at `http://3dsim.bioinfo.cnio.es`. The most important features of this web server are displayed in Figure 4.2.

3DSim can be queried in many different ways, although the simplest input is a CATH superfamily identifier to retrieve information on mapped mutations. Accordingly, the user can either manually introduce the desired superfamily identifier in the provided form or the complete list of superfamilies in CATH can be browsed to access the information (Figure 4.2, panel A). The database can be searched with UniProt accession numbers or CATH domain identifiers.

Once the CATH superfamily of interest is chosen, relevant information is displayed along with the list of CATH domains for which information is available regarding the mutations in SAAPdb. The number of pathogenic mutations is also reported for any CATH domain (Figure 4.2, panel B).

After a CATH domain of interest has been selected, 3DSim loads the main page for that specific domain. On this page, the server displays both an interactive Jmol plug-in that shows the mutations projected onto the 3D structure of the representative CATH domain and a 'mutation information table' (Figure 4.2, panel C) that contains all the information available for that given domain. This information includes the mutations available, the position of the mutations in the sequence and structure, their pathogenicity, and the similarity (BLAST sequence identity) between the sequences in Gene3D and the representative CATH domain sequence. Importantly, this similarity index provides the user with clues to the reliability of the homology-based transfer of mutations from the sequences in Gene3D to the structures in CATH. As a rule of thumb the stronger the similarity the more reliable the mutation transfer. By default, the server rejects mutations transferred from sequences with a BLAST sequence identity less than 20%, although due to the interactive approach of the server the user can establish more stringent constraints depending on the case under study.

3DSim also provides links to several external annotation sources where additional information about the mutations, proteins and structures can be obtained, including CATH [121], Gene3D [120], SAAPdb [74], Modbase [144], PDBsum [145] and UniProt [72]. Among these, the information SAAPdb may provide about the structural consequences of mutations is particularly interesting. Indeed, in some cases this information can help to understand the pathogenic characteristic of the mutations and provide an insight into the protein's function.

### 4.1.3 Access to the information: web services

Although punctual access to the information contained in the server is sufficient for most users, recursive programmatic access to the information is often necessary in genome-wide studies. To allow remote access to the resource, we developed nine SOAP web services that permit users to retrieve:

1. All known mutations for a given CATH domain grouped by a UniProt identifier.
2. The total number of mutations in a CATH domain.
3. All the CATH domains associated with a given UniProt identifier.
4. The amino acid sequence of a given CATH domain.

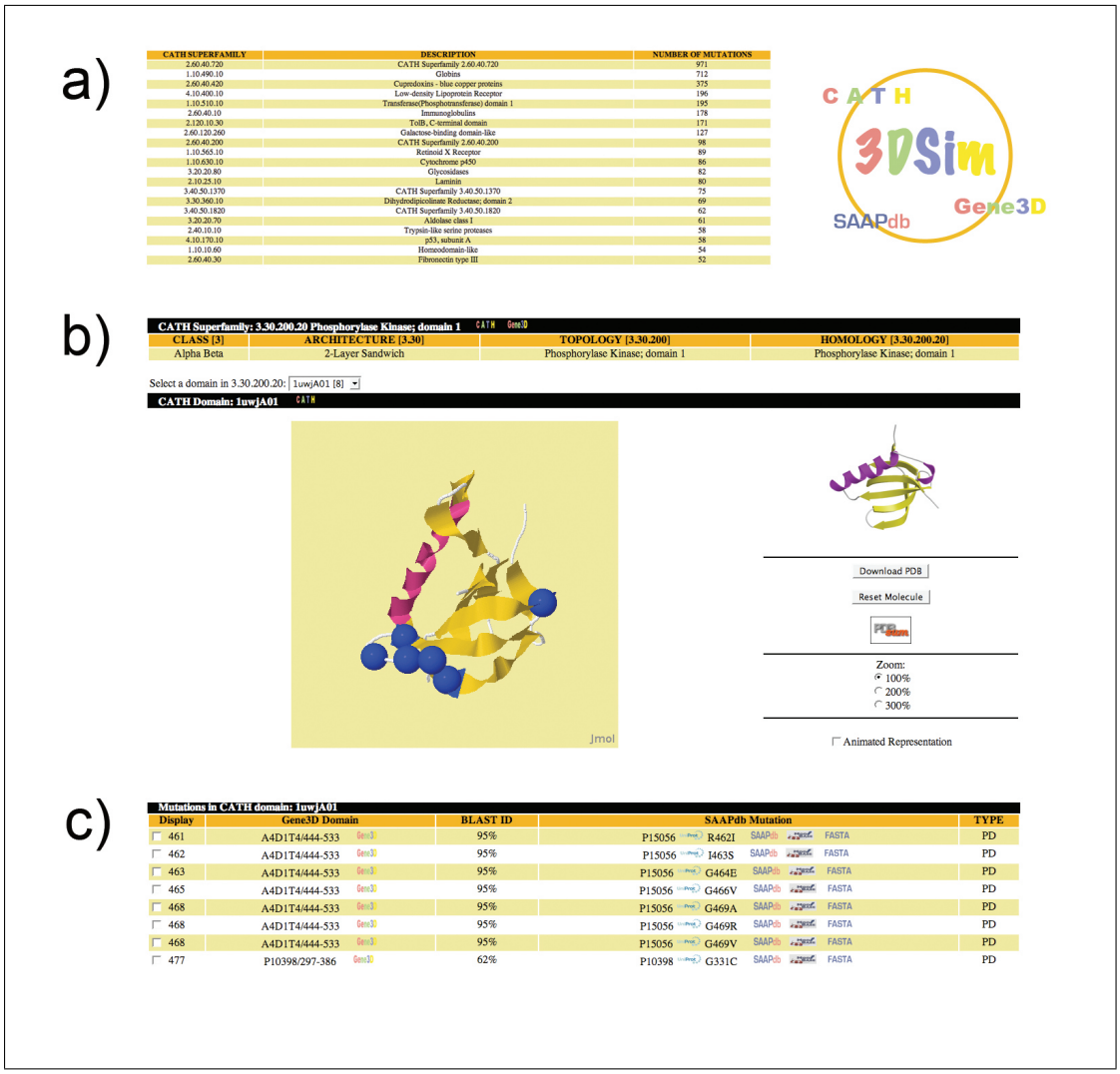



**Figure 4.2:** Schematic summary of the capabilities of 3DSim [143].
Panel A: Browsable list of superfamilies for which mutations exist.
Panel B: Example of a structure displaying known pathogenic mutations.
Panel C: An explanatory table that displays relevant information about the mutations in the chosen CATH domain.

5. All CATH domains in a CATH superfamily.
6. The superfamily to which a given CATH domain belongs.
7. The description and representative structure of a CATH superfamily.
8. All the mutations in SAAPdb for a given UniProt accession number.
9. The total number of mutations in SAAPdb for a given UniProt accession number.

These services were designed to facilitate the construction of elaborated computational pipelines. For example, a user starting from the UniProt accession number of a particular protein could retrieve a list of all the domains found in that protein and by chaining together four web service calls, the CATH superfamilies each domain belongs to, all the other domains that usually accompany the domain, and all the known mutations in the related domains can be found.

### 4.1.4 Example of the capabilities of 3DSim: the protein kinase superfamily

The protein kinase superfamily is subdivided into two different CATH superfamilies, each corresponding to a different structural lobe (Figure 4.3). The C-terminal lobe of the kinase is represented by the 1.10.510.10 CATH superfamily, which corresponds to the phosphotransferase domain I homology group in CATH, and it includes 73 different representative subdomains harboring mutations. Of them, subdomain 1rw8A02 (Figure 4.3, panel A) contains the largest number of mutations, 29 mutations that correspond to 22 different residues in the protein structure. Of these 29 mutations, only 3 come from a sequence, namely P36897, which maps directly onto the domain. The remaining 26 correspond to homologous sequences with 40-82% sequence identity according to Gene3D. The ability to increase the number of mutations in a given structure by homology-based transfer is probably the most important added value of the server.

Interestingly, the pathogenic mutations reported in 3DSim for the human TGF-$\beta$ receptor type I (M318R, D400G and R487P) have already been associated with Loeys-Dietz syndrome type 1A, an aortic aneurysm syndrome with widespread systemic involvement [146]. This disorder is characterized by arterial tortuosity, aneurysms, craniosynostosis, hypertelorism and bifid uvula (cleft palate). Exotropy, micrognathia, retrognathia, structural brain abnormalities, intellectual deficit, congenital heart disease, translucent skin, joint hyperlaxity and aneurysm with dissection throughout the arterial tree have also been reported.

The structure of TGF-$\beta$ receptor type I (1rw8) can be seen along with the pathogenic deviations from SAAPdb (Figure 4.3, panel A), and a similar image can be obtained directly from our server and is one of the main features available to analyze the distribution of mutations within structures. This figure reveals that pathogenic deviations tend to cluster around important structural features, such as the catalytic loop or the substrate-binding groove.

The N-terminal lobe of the protein kinase superfamily is represented by the CATH superfamily 3.30.200.20, which contains 21 different subdomains with at least one reported mutation. The one with most mutations is 1uwjA01 (Figure 4.3, panel B), accounting for 8 pathogenic deviations in 6 different residues. Interestingly, none of these mutations corresponds directly to the structure of the representative protein but rather, to structurally similar proteins in the same subdomain classification that have been transferred by our algorithm. The mutations inherited at 95% identity from the human serine/threonine-protein kinase B-raf are particularly noteworthy, a well-known proto-oncogene involved in the transduction

of mitogenic signals from the cell membrane to the nucleus. Mutations in this gene cause cardiofaciocutaneous syndrome and a wide number of diseases such as lung cancer, colon cancer, melanoma and several types of cancers affecting the immune system, including non-Hodgkin lymphoma. As a matter of fact, the mutations reported in our server have been linked to colorectal cancer (R462I, I463S and G464E) [147], lung cancer (G466V) [148] and non-Hodgkin lymphoma (G469A, G469R/V) [149].

From the structure of B-raf and the pathogenic mutations reported in the database (Figure 4.3, panel B), it is evident that the mutations are located near an important structural element, such as the P-loop, which is involved in ATP binding. In fact, some of the mutations target Gly-464, Gly-466 and Gly-469, the highly conserved amino acids in the glycine-rich motif GxGxxG that interacts with the $\beta$- and $\gamma$-phosphates of ATP.

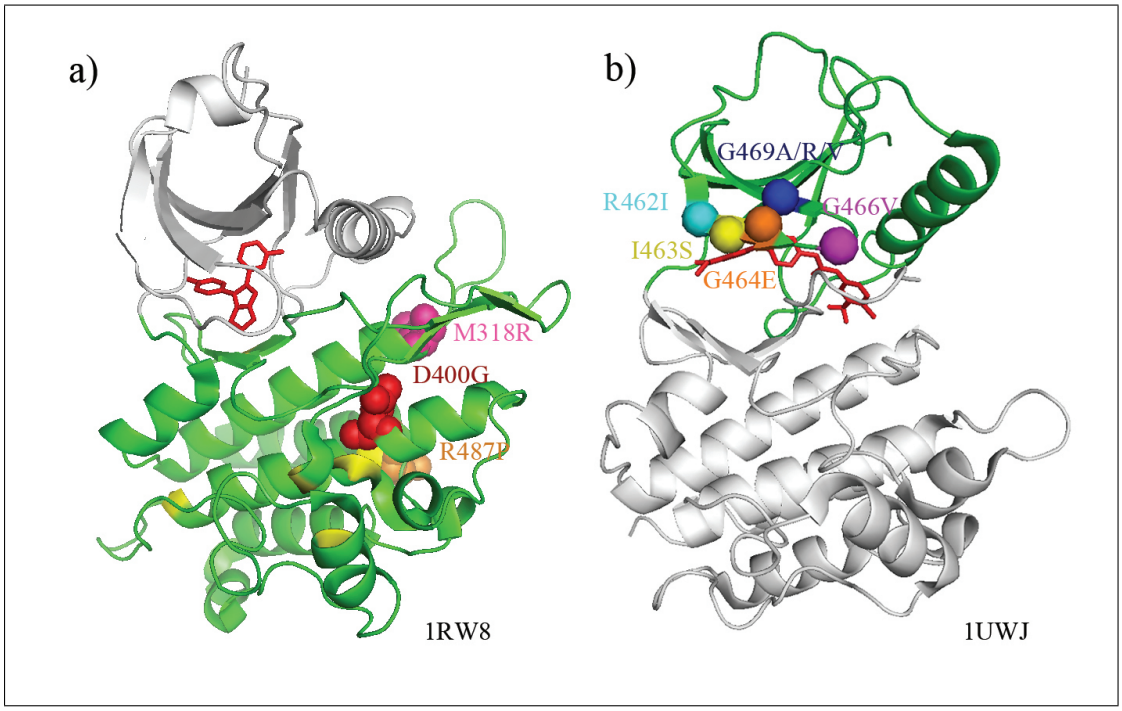



**Figure 4.3:** Examples of mutations mapped onto representative members of the protein kinase superfamily.
Panel A: Pathogenic mutations mapped onto the structure of the TGF-$\beta$ receptor type I (1rw8A). Mutations transferred from other proteins with the same structural representative have been highlighted in yellow, whereas the pathogenic mutations occurring directly in the structure of the TGF-$\beta$ receptor have been highlighted as follows: M318R in violet, D400G in red and R487P in orange.
Panel B: Pathogenic mutations in B-raf (1uwjA). Cancer-associated mutations in the P-loop have been highlighted: R462I in cyan, I463S in yellow, G464E in orange, G466V in magenta and G469A/R/V in blue.

## 4.2 Automatic literature-mining

Mutations discovered from either high-throughput genome wide studies or from detailed studies of specific kinases in various biological systems are often recorded in state-of-the-art databases, such as the SwissProt Variant Pages [71], COSMIC [27] or KinMutBase [75]. However, large-scale human variation studies generate a vast amount of information that is not properly resourced to store, annotate and curate. Most of the existing manually-curated mutation annotation resources rely on reading a subset of all published articles, while existing automated systems are mainly based on (subsets of) PubMed abstracts or small (but broader than the manual ones) collections of full-text articles [150]. Please refer to the Introduction for a review of these resources.

Therefore, many mutations are not stored in databases. When annotations are lacking, it is not easy to recover functional information even when the experimental results have been published. Whereas manual inspection of the literature, curation and annotation of mutations is possible for very specific singular cases, it becomes practically unfeasible for large sets of proteins and/or documents.

Thus, we tackled the difficulties in obtaining generalized information about mutations in the protein kinase superfamily, which represents a large superfamily with a growing number of published reports. Accordingly, we needed a system capable of recovering kinase mutations from the literature in a fully automated manner. We implemented a pipeline that integrated article retrieval, the detection of mutations mentioned in the literature, and a final validation of the mutations linked to their corresponding protein sources [150]. Although the pipeline is focused on mutations within the protein kinase domain, it constitutes a prototype that might be easily applied to any other superfamily. This pipeline (Figure 4.4) includes the construction of a kinase-relevant article collection that considers both abstracts and full-text articles from PubMed, the detection of mutation mentions, the predictive classification of mutations into induced or natural, the linking of mutations to corresponding protein sequences, and comparisons to existing databases.

### 4.2.1 Mutation mention extraction and disambiguation

Our mutation extraction pipeline was applied to two different document sets, one containing the whole collection of PubMed abstracts and the other, a collection of 19,404 full-text articles. The full-text articles were automatically downloaded using an in-house retrieval system [79] following three different criteria:

1. Relevance of the abstract: information contained in the corresponding abstracts such as the mention of mutations, mention of human kinases and a combination of keywords (including 'human kinase mutation').

2. A priori relevance of the articles: extracting all the PubMed references for human kinases contained in multiple databases (e.g., SwissProt, MINT and IntAct).

3. Relevance of the journal: based on analyzing a fraction of mutation-mentioning abstracts of each journal and prioritizing a set of journals (and thus their articles) to retrieve their full-text articles. These journals included the *American Journal of Human Genetics, European Journal of Human Genetics, Human Genetics, Human Mutation and Human Molecular Genetics.*

Both abstracts and full-text articles were then pre-processed by applying an in-house rule-based sentence boundary detection system that we optimized for PubMed abstracts [79].

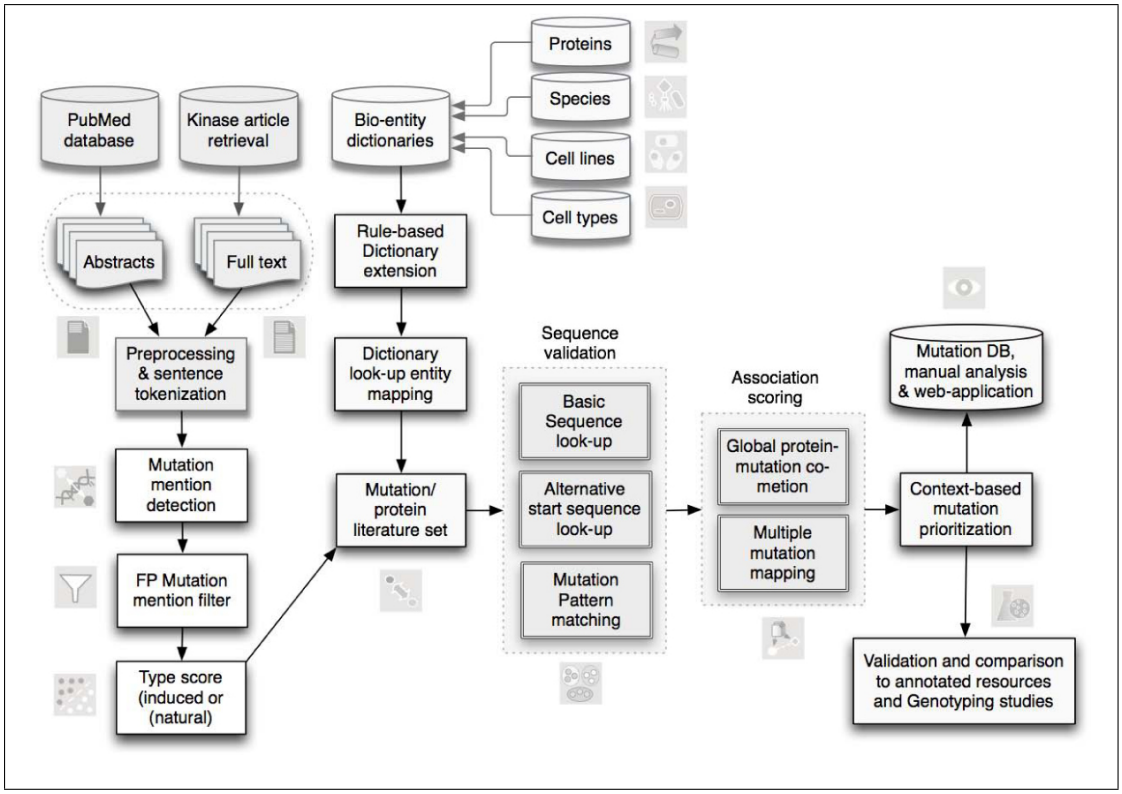



**Figure 4.4:** Flow chart of the literature-mining pipeline for mutation extraction presented, giving an overview of the different processing steps to extract interesting mutations in the human kinome.

For the initial extraction of single amino acid substitutions we used MutationFinder [90], a software for point mutation literature-mining based on regular expressions and patterns. Its main function is to detect the mention of mutations in a given set of manuscripts and it relies on language expressions used to describe mutation events. This system is very competitive for recall and precision when compared to other strategies [84], and it has been evaluated using a manually-generated gold standard collection of abstracts. In addition to the raw results provided by MutationFinder, we designed an approach to target mutation pattern sense disambiguation and the filtering of mentions that do not correspond to protein mutations. To determine whether a candidate mutation mention corresponds to a mutation or an artifact, we designed a module that filters false positive mentions through a combination of named entity recognition, dictionary look-up and rule-based methods. Most false positive mutation mentions corresponded to one of the following three semantic types:

1. Cell lines or cell types. Several frequently mentioned cell lines resembled mutation mentions, including the human glioblastoma cell line T98G, the T-cell line M14T, the adrenocortical cell line H295R, and other commonly used cell lines such as T47D or T24C.

2. Taxonomic entities. Certain taxa, especially bacterial strains, cloning vectors and some mouse models displayed names that were similar to single letter mutation mentions e.g., *Escherichia coli* K12S, *Actinomyces viscosus* T14V, *Pneumocystis pneumoniae* R36A, *Actinomyces naeslundii* T14V, *Mycoplasma sp.* G145T and the yeast strain S288C. Clone identifiers (e.g., W12I and W12E) or plasmids (e.g., *E. coli* plasmids P15A) can also cause false positive mutation identification. A special case of ambiguous mutation mentions are

transgenic mouse models like G93A transgenic mice. Although they are a strain expressing a G93A mutant of the human SOD1 protein, these mice are usually mentioned in the literature as the strain rather than as a reference to the particular mutation.

3. Protein, gene and drug names. Several protein names match the patterns used to identify mutations from the literature. Although some of them correspond to human proteins, like S100D and S100E, a considerable fraction are viral gene names (e.g., *vaccinia* viral A10L, *variola* viral A11L or the poxvirus protein A52R). We found some additional cases of wrongly tagged mutations that could be classified as drugs or compounds (e.g., the antibiotic A83586C, the immunogen A27L or the antifungal antibiotic A9145C). To determine the semantic class of a given mutation, we explored the use of knowledge-based methods relying on machine-readable dictionaries (MRDs) for sense disambiguation based on local context analysis.

### 4.2.2 Linking mutation mentions to human kinase sequences

Linking mutations that appear in the literature to their corresponding protein sequences and database records is crucial to characterize the putative structural implication of the mutations. This also allows direct comparison of the mutations retrieved to manually-curated functional annotation of protein mutations contained in dedicated databases, as well as the integration of the literature mentions into large-scale experimental genotyping studies. Here, we focused on the association of the literature-extracted mutation mentions with human protein kinases. More specifically, we restricted our analysis to mutations occurring within the protein kinase domain defined in KinBase [40].

To link mutation mentions and the sequences of human kinases, we assumed that the corresponding protein names were co-mentioned in the articles. After extracting mutation mentions from PubMed abstracts and a large collection of full-text articles, these two data sets were explored for mentions of human protein kinases. We applied a dictionary look-up approach to detect any mentions of kinase proteins, similar to strategies that were successfully used in the gene normalization [1] task of BioCreative II [151]. To take into account inter- and intra-species protein name ambiguity, instead of using very strict protein-organism source co-mention criteria based on relative textual distances, we calculated two scores for each article reflecting: (1) the contextual similarity of the article to the SwissProt protein record; and (2), the overall association of the article with human species terms from the total set of tagged species terms. This high recall protein normalization scoring strategy was followed by a stricter sequence validation that detected links between the mutations and proteins by checking whether the mutation mentions could be confirmed by the presence of the residues at the precise sequence positions.

To enhance the recall of the basic sequence look-up validation method, we implemented five complementary mutation-sequence mapping strategies. These took into account errors resulting from the wrong detection of the actual directionality of the extracted mutation with respect to wild type and mutant residues, as well as inconsistencies and alternative sequence counting between the article and the database kinase sequence. The criteria are summarized as follows:

1. *Basic mutation to sequence position mapping.* Basic validation system as described previously where the algorithm looks for the wild type residue of an extracted mutation mention at the corresponding protein sequence position.

[1] In the text-mining domain normalization is the task of linking entities (i.e. protein names), to a common reference (i.e. a protein database entry)

2. *Alternative mutation directionality look-up.* To account for errors in the automatic extraction of the mutation mention directionality (i.e., a mutant residue is considered wild type by the pipeline), we examined whether the mutant residue could be matched to the corresponding sequence position.

3. *Relative mutation patterns.* We considered a sliding window algorithm that searches for a pattern of mutations in a sequence, in addition to the exact position co-occurrences within the sequence presented previously. The algorithm recursively scans each position in the sequence and searches for co-occurrence of the other mutations mentioned in the same abstract in positions relative to the start, which is used instead of the exact positions provided to consider the distance between all the mutations in sequence terms. Thus, this approach has to deal with the different ways the starting position of a protein can be defined, the most clear being the presence or absence of a signal peptide, but other examples may arise (e.g., sequencing errors or discrepancies, inclusion of promoter regions, etc...). Since finding a profile by chance is quite easy for trivial results (the easiest being patterns comprising just one mutation), a limitation in the complexity of the pattern was established, taking into consideration only patterns with at least 3 mutations at different sequence positions.

4. *Pro-peptides and mature protein mutation mapping.* To handle alternative residue counting when a signal peptide is present, we looked for the wild type residue in a map that considers the additional length of existing N-terminal signal peptides.

5. *Methionine cleavage start site counting.* We mapped a mutation by taking into account putative methionine cleavage and neglecting the N-terminal methionine.

A total of 567 triplets (i.e., article-mutation mention-protein associations) derived from the abstract corpus was validated by the basic wild type to position mapping. By applying the additional 4 matching strategies, we recovered 437 additional hits, corresponding to 43.53 % of the total set of sequence-validated protein-mutation pairs. This added up to a total of 1,004 triplets from 714 abstracts.

When the full-text corpus was considered, the total number of triplets detected by the basic mapping was 3,911, while another 3,917 triplets were recovered by the additional sequence mapping methods. This resulted in 7,828 triplets from 3,496 full-text articles. The average number of sequence-validated mutations in the protein kinase domain was 1.41 and 2.24 for each abstract and full-text article, respectively, implying that each paper often describes more than one mutation.

When we considered unique mutation-protein pairs irrespective of the number of times this pair was recovered from the literature, a total of 643 kinase domain mutations from 128 different protein kinases were extracted from PubMed abstracts. When considering the full-text collection, this number increased considerably and 6,970 mutation-protein pairs were obtained from 325 protein kinases. Therefore, using full-text articles significantly increased the mutations recovered (more than 10 times more mutation-protein pairs when compared to those obtained from the abstracts) and also increased the recall of proteins for which mutations had been extracted (more than doubling the initial number derived from abstracts alone). The increased recall with full-text papers justified the computational effort required to retrieve and pre-process them.

### 4.2.3 Manual validation of a representative subset of mutation mentions

To determine to what extent the information extracted could be trusted as a source of potentially missing mutations that still needed to be annotated in the databases, we manually validated a random sample of 100 mutation mentions from PubMed abstracts.

We found (Figure 4.5) that for 23% of the mutations, there was a confirmed record in at least one of the analyzed knowledge bases (we will discuss this aspect more thoroughly below). Moreover, 41% of the results corresponded to correct protein-mutation assignments obtained by our extraction pipeline that were missing in the repositories studied. In addition, 8% of the mutations were examples of mentions from orthologues of human proteins having the same amino acid at the specified position. These results can also be considered positive hits since they mostly correspond to animal models for the indirect analysis of human kinases. Interestingly, a small proportion (2%) of the records corresponded to mutations too ambiguous even for human curators and therefore, they were inherently unpredictable for the automated methods.

We estimate that around three quarters (72%) of the mutations that we extracted with our automated text-mining pipeline corresponded to correct mutation mentions in the literature. These correct predictions integrate mutation mentions that correspond to previously annotated mutations, and reliable novel mutations published in the literature but not recorded in the knowledge bases for both humans and closely orthologous models.

### 4.2.4 Evaluation of the mutation extraction pipeline by comparison to existing repositories of experimentally curated data

We assessed whether the mutations recovered from the literature by our system were already present in commonly used databases or if they were newly recovered. Accordingly, we studied the overlap between the mutations in the protein kinase domain recorded in some representative state-of-the-art knowledge bases and the mutation mentions automatically retrieved by our literature-mining pipeline.

Moreover, we wanted to assess how many mutations we could recover from a combined dataset of mutations from all the studied repositories in order to determine our extraction pipeline's coverage of existing knowledge. We built a non-redundant set with 1,265 mutations in 317 different kinases, which also highlighted that the different knowledge bases were unevenly represented depending on the purpose of each database. The weight of each database is reported in Table 4.1, where the overlap between the different databases is assessed.

Out of the 1,265 mutations in the combined dataset, 148 (11.70%) were found by our automated pipeline when the PubMed abstracts were scanned. By contrast, 354 (27.98%) mutations were recovered when full-text articles were considered and 399 (31.54%) when the combined abstract/full-text dataset was used. The recall of mutations from the databases is a way to measure the capacity of the methodology and it provides additional information in the form of descriptive sentences about the recorded mutations, which clearly justifies the computational effort required. This aspect of the extraction pipeline will be discussed thoroughly below.

Interestingly, only a small fraction of the mutations detected in the datasets of high-throughput genotyping studies (COSMIC [27] and Greenman/Wood [14, 16]) corresponded to previously identified mutations mentioned in the literature. This is consistent with the fact that mutations from high-throughput projects are generally not studied in detail and they are

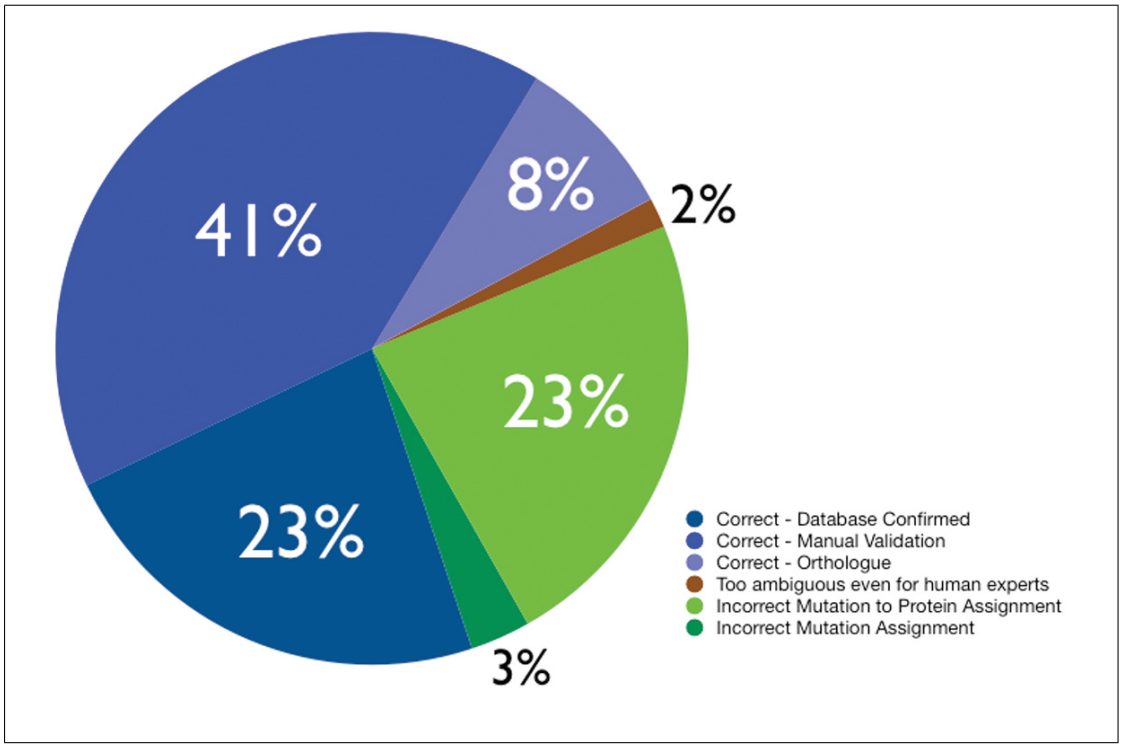



**Figure 4.5:** Success estimate of the extraction pipeline by expert human manual validation. These percentages were calculated by a manual sampling and validation protocol carried out on 100 abstracts. The categories are described as follows:
*Correct - Database Confirmed*: These are the mutations that have already been found in at least one of the mutation repositories analyzed (Uniprot, SAAPdb, COSMIC, KinMutBase or Greenman/Wood).
*Correct - Manual Validation*: This subset corresponds to the mutation-protein pairs that were found to be correct after manual validation of 100 abstracts.
*Correct - Orthologue*: This subset corresponds to the cases where mapping is confirmed by manual validation and the mutation is mapped to a non-human orthologue.
*Incorrect mutation to protein assignment*: This corresponds to the cases where at least 2 proteins in the same text share the same amino acid at the mutated position and the algorithm confuses the pairing.
*Incorrect mutation assignment*: These are cases where the mutation is not properly identified.
*Too ambiguous even for human experts*: These are cases that lack enough supporting information even for human curators.

therefore not reported in the literature.

The results for the neutral subset of SAAPdb [74] are particularly striking. Except for a couple of cases, we did not find references to mutations from this set in the literature, regardless of the type of collection analyzed. This is probably because a considerable fraction of these mutations have a neutral phenotype and thus, they are not mentioned further in the literature. Another interesting observation was that contrasting results were obtained for natural variants and induced mutations in the SwissProt Variant database [84]. We observed differences in the percentage overlap of mutations recovered from abstracts and full-text articles when the two types of mutations were considered individually. This result suggests that experimentally-induced mutations annotated in SwissProt are usually not mentioned in the abstracts but rather, they only appear in the full-text articles, while the converse is true for natural variants. This agrees with the underlying idea of natural variants being primary discoveries in their respective papers, whereas induced mutations are usually reported for experiments whose goal is not to report novel mutations but biochemically characterize the wild type residues.

There were mutation records that we could not find any evidence of in the literature and among the reasons why our system could not detect them, the most plausible are:

1. Lack of accessibility to full-text articles or additional materials.
2. General limitations in terms of recall of the mutation mention extraction methods.
3. Limitations in protein normalization.
4. Limitations in the association of mutations with corresponding sequences.

| **Knowledge base** | **Muts.** | **Weight** | **Abstract** | **Full-text** | **Combined** |
|---|---|---|---|---|---|
| SwissProt; all | 710 | 56.13% | 134 [18.87%] | 328 [46.20%] | 365 [51.41%] |
| SwissProt; natural variants | 459 | 36.28% | 99 [21.57%] | 196 [42.70%] | 230 [50.11%] |
| SwissProt; mutagenesis | 251 | 19.84% | 35 [13.94%] | 132 [52.59%] | 135 [53.78%] |
| SAAPdb; all | 610 | 48.22% | 65 [10.66%] | 106 [17.38%] | 125 [20.49%] |
| SAAPdb; pathogenic deviations | 323 | 25.53% | 64 [19.81%] | 105 [32.51%] | 123 [38.08%] |
| SAAPdb; neutral SNPs | 287 | 22.69% | 1 [0.35%] | 1 [0.35%] | 2 [0.70%] |
| Greenman/Wood; all | 254 | 20.08% | 4 [1.57%] | 12 [4.72%] | 13 [5.12%] |
| Greenman/Wood; driver | 119 | 9.04% | 3 [2.52%] | 9 [7.56%] | 9 [7.56%] |
| Greenman/Wood; passenger | 135 | 10.67% | 1 [0.74%] | 3 [2.22%] | 4 [2.96%] |
| COSMIC | 200 | 15.81% | 4 [2.00%] | 11 [5.50%] | 12 [6.00%] |
| KinMutBase | 83 | 6.56% | 32 [38.55%] | 32 [38.55%] | 43 [51.81%] |
| All databases | 1265 | - | 148 [11.70%] | 354 [27.98%] | 399 [31.54%] |

**Table 4.1:** Coverage in the existing knowledge bases of the mutations extracted from the literature. The percentages in brackets show the fraction of each database recovered by our text-mining pipeline.

### 4.2.5 Phylogenetic distribution of the extracted mutations

We analyzed putative biases in the distribution of the mutation mentions by superimposing the topology of the protein kinase superfamily defined by KinBase [40] onto the results presented above (please refer to the Introduction for a more detailed description of the different kinase groups). There are large differences in the total number of mutations in each of the clades,

with more than half of the mutations in either the TK or CMGC groups (the distribution of mutations across the different kinase groups is shown in Figure 4.6, panel A). Interestingly, our system could extract mutations from all the groups using either PubMed abstracts or full-text articles.

The normalized distribution of mutations in the different protein kinase domains defined by KinBase was established when either abstracts or full-text articles were used (Figure 4.6, panel B). It is clear that no matter which dataset was used, abstract or full-text articles, the distributions were very similar and independent of the absolute number of each dataset.

### 4.2.6 Location of the extracted mutation mentions in the protein kinase domain

Intuitively, relevant regions in the protein structure would accumulate more mutation mentions and conversely, the number of times a particular mutation is mentioned would indicate the relevance of that mutation for protein function. Thus, we considered the mutation density distribution within a consensus protein kinase domain model (Figure 4.7). Although mutations were scattered all around the consensus structure of the kinase domain, a higher mutation density was found close to functionally relevant elements, such as the ATP-binding pocket or the DFG motif in the activation loop. In fact, the ATP-binding Lys-64 is associated with most mutations, with a total of 65 mutations, followed by several regions in the activation segment with up to 39 mutations per residue. By contrast, regions with a low mutation density are not functionally relevant.

### 4.2.7 The other side of the coin: from mutations to the literature

We described earlier the importance of our system in retrieving mutations from the literature that are not reported in the databases. A second application of the method that is at least equally importantly is its use as an information repository. The added value of our system is the vast amount of manuscript references and sentences that describe the experimental conditions and the implications of the mutations. The latter constitute an important source of information to help assess the pathogenicity of mutations, and in the best possible scenario, the biochemical mechanism and/or phenotypic consequences.

Mutations in the epidermal growth factor receptor (EGFR) provide a working example of such a utility. EGFR is a protein kinase involved in controlling cell growth and differentiation, which has been linked to breast cancer development. EGFR ligation elicits dimerization, internalization of the binary complex, induction of tyrosine kinase activity, stimulation of cell DNA synthesis and cell proliferation. Several well-known mutations in this protein are present in the current state-of-the-art databases: SwissProt [84], COSMIC [27], Greenman/Wood [14,16], KinMutBase [75], and SAAPdb [74]. In some cases, the contribution of these mutations to disease has been studied and annotated in the corresponding databases. For instance, the somatic mutations G719S, L858R and T790M have already been associated with lung cancer [14,152].

Our system recalled a further 32 novel mutation mentions from the literature that have not been reported in the databases. To better understand the effect of these new mutations, our approach provided contextual information to help interpret the role played by the mutations. For example, in the case of Y845F (transformed to Y869F due to the presence of a signal peptide) we found the following sentences: *'Furthermore, transient expression of a Y845F variant EGFR in murine fibroblasts resulted in an ablation of EGF-induced DNA synthesis to non-stimulated levels.'* (PubMed:10075741); *'Stably transfected B82L cells with a point*

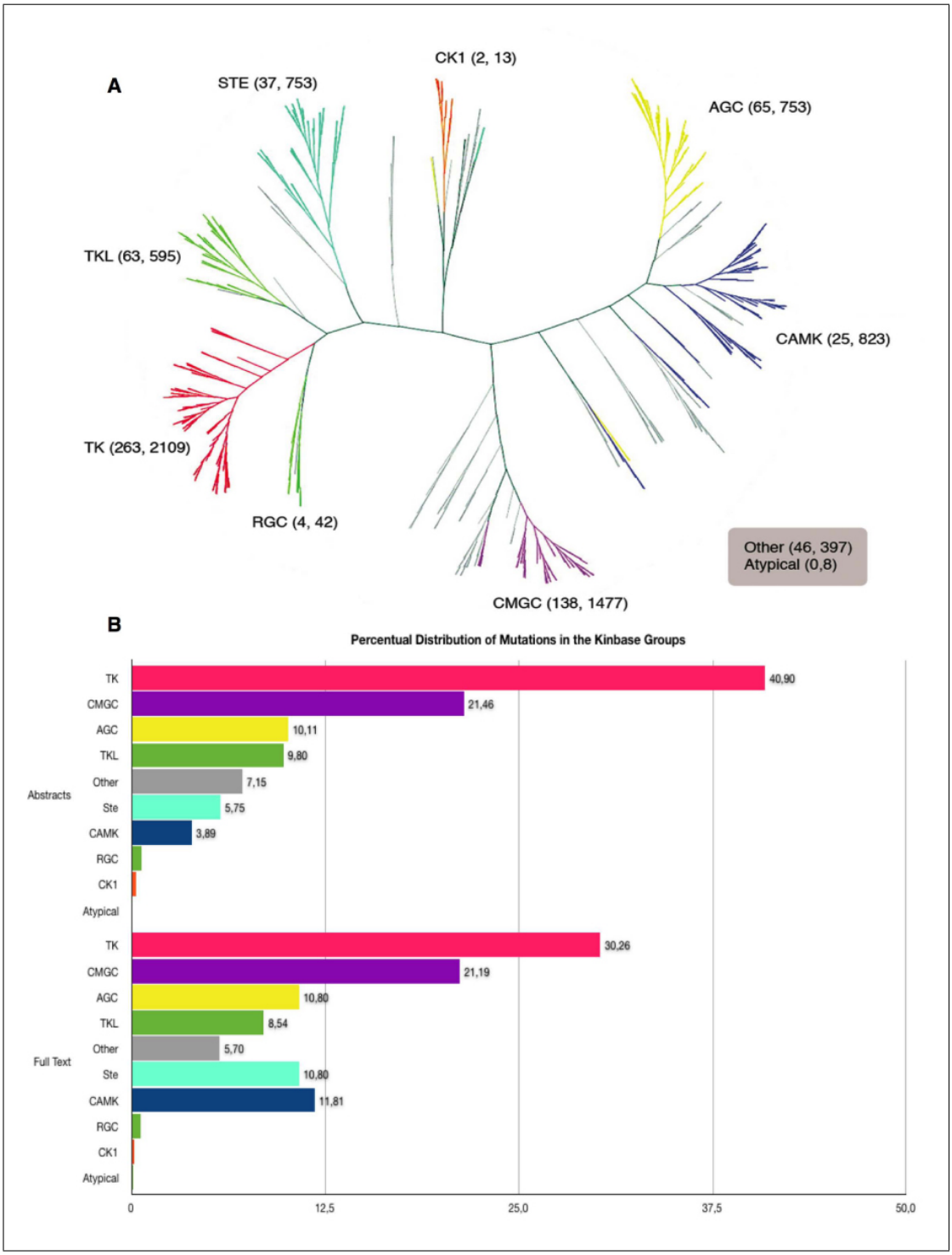



**Figure 4.6:** Distribution of literature-extracted mutations in the groups defined by KinBase [40].
Panel A: Number of mutations from the literature housed in the different protein kinase domain groups defined by KinBase when either abstracts or full-text articles are used.
Panel B: Normalized distribution of mutations in the topology of the protein kinase superfamily when either the abstract or the full-text collection is considered.

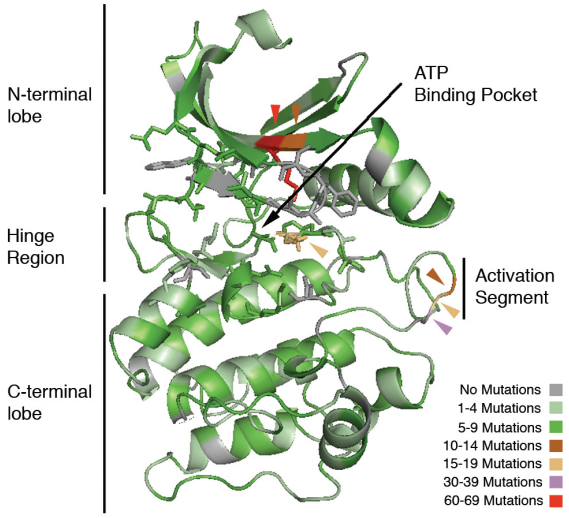



**Figure 4.7:** Density of mutations extracted from the literature within the structure of a consensus protein kinase domain model. The ATP-binding pocket is represented by sticks. The residue with the highest density of mutations is the ATP-binding Lys-64 (red sticks), with a total of 65 mentions, followed by residues in the activation segment (up to 39 mutations per residue) and others in the ATP-binding pocket. The DFG motif (activation segment essential for kinase function) accumulates many mutations. The light brown asparagine (central part of the figure) in the inter-lobe region has more than 10 mutations.

*mutation of the EGFR at Tyr-845 (B82L-Y845F) exhibited only basal Ras activity following exposure to* $Zn^{2+}$*'* (PubMed:11983694); and *'In contrast, LPA-elicited DNA synthesis and migration were augmented in cells expressing EGFR, EGFR(K721A), or EGFR(Y845F), but not EGFR(Y5F), although the PDGF responses were indistinguishable'* (PubMed:15364923). The information retrieved suggests that Tyr-845 is involved in DNA synthesis provoked by EGF binding to the receptor.

Our system also retrieved functionally neutral results that are often discarded and not stored in the databases, despite containing useful information for the contextual interpretation of the involvement of specific residues in protein function: *'Unexpectedly, the Y845F mutant EGFR was found to retain its full kinase activity and its ability to activate the adapter protein SHC and extracellular signal-regulated kinase ERK2 in response to EGF, demonstrating that the mitogenic pathway involving phosphorylation of Y845 is independent of ERK2-activation'* (PubMed:9990038). The structural model of this protein and a summary of the information gathered from the literature by our system regarding specific residues and mutations are represented in Figure 4.8.

## 4.3 Characterization of the residues that disrupt protein kinase function

In the examples analyzed above, we found that pathogenic mutations targeted structurally and functionally important residues, while neutral polymorphisms were scattered throughout the protein (Figure 4.9). To further explore this issue, we assessed whether there was a preferential distribution of the amino acids according to their pathogenicity and to what extent the pathogenicity of the mutations was explained by the relevance of the residues harboring them. In this study we focused on protein kinases, a well-studied set of proteins associated with cancer onset and progression.

We mapped the different types of mutations onto a consensus structural model representing the whole protein kinase superfamily and analyzed their distribution at evolutionary conserved positions and known functional regions (buried, functional and conserved). Two

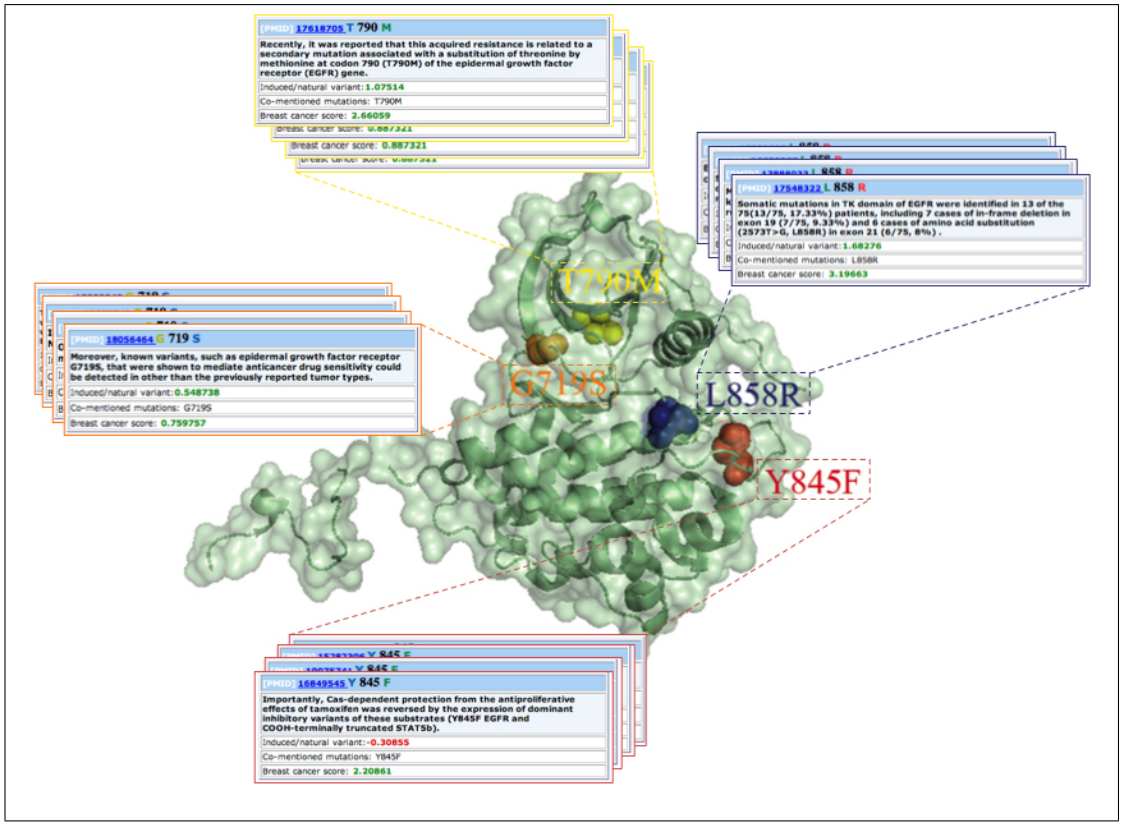



**Figure 4.8:** Mock-up of the structural model of the epidermal growth factor receptor (EGFR, pdb 1m17) together with a summary of the residue and mutation information gathered from the literature by our automatic literature-mining pipeline.

different groups of datasets were used for the comparison, one with disease-associated mutations and the other with neutral polymorphisms. Driver somatic mutations (those predicted to be involved in cancer onset) were compared to passenger ones (supposedly neutral) [13, 14, 16], while pathogenic germline deviations (mutations with structural evidence linking them to disease) from SAAPdb [74] were compared with neutral polymorphisms.

In the Introduction, we elaborated on the differences between these two separate groups of mutations. Germline mutations are inherited and therefore, they are present in every cell of the individual, whereas somatic mutations are acquired after conception, mainly induced by external agents or duplication errors. Consequently, somatic mutations do not occur in all the cells of an individual and they are not transmitted to the offspring.

Irrespective of the dataset, the significance of the proximity of different sets of mutations to specific areas of the protein was assessed through the $Xd$ measure [133]. We chose this weighted measure of distance distributions to prioritize the differences in regions closer to the regions studied (for example, binding sites) as opposed to differences in the distribution of residues far from the regions of interest. The $Xd$ value has been used as a standard to measure the difference between the distribution of predicted residues in the context of the CASP challenge [134, 135, 153]. For full details see the Methods section.

To provide a single measure that summarizes the behavior of the whole superfamily

while accounting for a reasonable number of mutations capable of achieving reliable statistical significance, we transferred the mutations from each specific protein kinase onto a consensus model that we considered a good proxy for the common structural features of the superfamily (Figure 4.9). The protein kinase superfamily is amenable for this type of approach due to the structural similarities among its members [154, 155].

### 4.3.1 A consensus model of the protein kinase superfamily

A consensus model of the basic structure of the kinase domain was constructed, representing the average structure of a large number of kinases in the human kinome and capturing the common characteristics of these proteins in a single structure. To build the model, we first selected MAP3K1 as a standard representative sequence of the family from a manually-curated multiple sequence alignment of the human kinome constructed with MUSCLE [123]. A model of the sequence selected was built with Modeller [156], which assembled all the closely related PDB template structures returned from a BLAST search [122]. The consensus model and some of the most important functional regions are shown in Figure 4.9.

### 4.3.2 Distribution of somatic driver and passenger mutations

In this experiment, the datasets used corresponded to the somatic mutations discovered in cancer resequencing projects [14, 16] and that were located within the protein kinase domain. The pathogenic dataset corresponded to the somatic mutations classified as 'drivers' (those that are more probably disease-associated), while the neutral dataset consisted of 'passenger' mutations (those not thought to be pathogenic).

We assessed the significance of the proximity of different sets of mutations to specific areas of the protein using the harmonic deviation, $Xd$ [133]. This was performed to prioritize the differences in the regions near to the studied regions (for example, binding sites) over the differences in the distribution of residues in positions far from the regions of interest. To compare the two distributions, $\Delta Xd$ was calculated as the difference of $Xd_{passengers}$ and $Xd_{drivers}$. Greater positive values indicated that passengers localize closer than drivers (on average) to the residues annotated with the feature studied, while greater negative values indicated that drivers co-localize with the regions under study. For explanatory purposes, an arbitrary threshold of 0.5 was chosen to consider the two distributions of distances to be sufficiently different. The results of these analyses are displayed in Table 4.2.

| | **Mean Dist. drivers (Å)** | **Mean Dist. passengers (Å)** | $Xd_{driver}$ | $Xd_{pass}$ | $\Delta Xd$ |
|---|---|---|---|---|---|
| Seq. conservation, Shannon | 7.26 | 7.43 | 1.27 | 0.04 | -1.23 |
| Seq. conservation, AL2CO | 8.46 | 8.51 | 1.42 | 0.54 | -0.88 |
| Accessibility | 3.88 | 3.77 | -0.88 | -0.87 | -0.01 |
| Catalytic site, FireDB [127] | 11.36 | 10.80 | 0.56 | -0.39 | -0.95 |
| Catalytic site, Knight [46] | 14.34 | 13.85 | 0.55 | 0.00 | -0.55 |
| Tree-determinants | 6.50 | 6.71 | 1.25 | -0.30 | -1.55 |

**Table 4.2:** Distribution of driver and passenger somatic mutations in regions that are evolutionary conserved, that display structural conservation or that retain functionality.

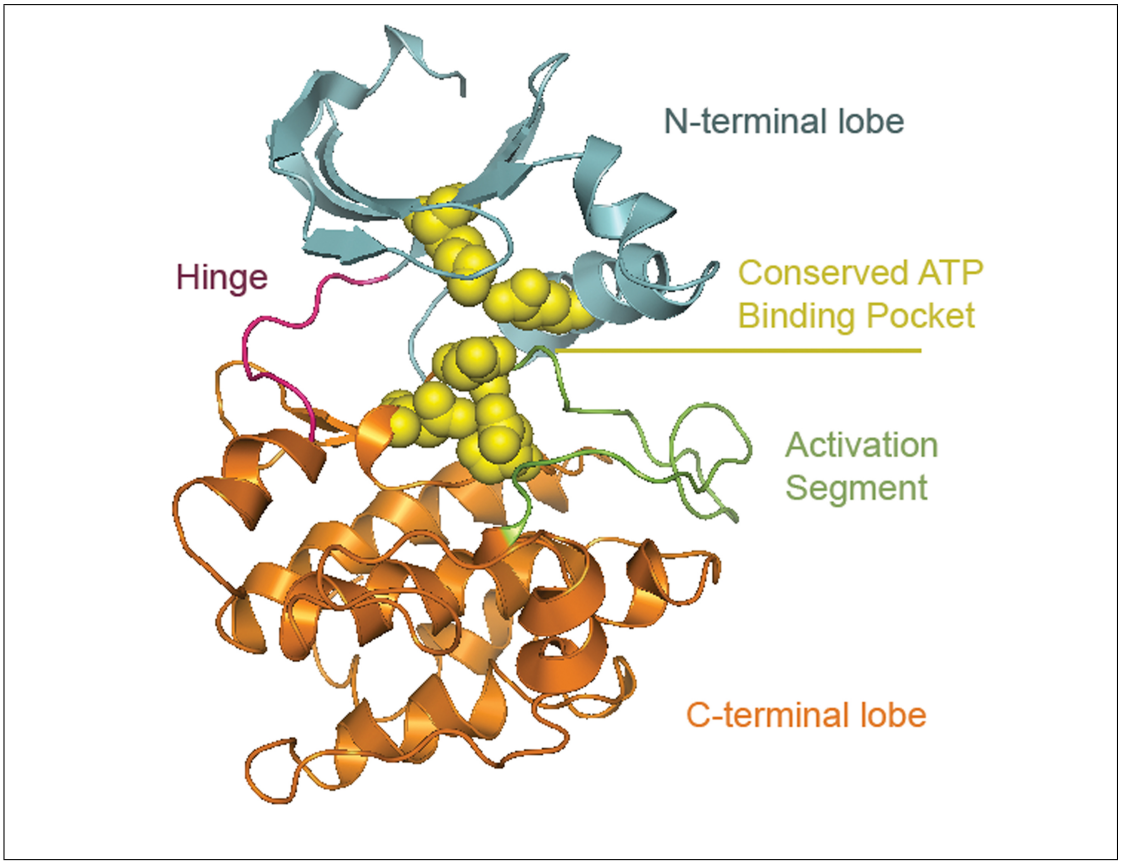



**Figure 4.9:** Our model structure of a human protein kinase domain based on MAP3K1. The model shows the basic two-lobe kinase fold, with the N- and C-terminal (cyan and orange, respectively) lobes joined by a hinge region (magenta). Recognition of the substrate protein is mainly through interaction with the activation segment (green), a region in the C-terminal lobe. ATP binds at a site between the two lobes, where five highly conserved residues guide the positioning of the molecule: K74, E96, D171, N176 and D190 (yellow, numbers corresponding to positions in the generated structural model). The substrate-binding groove is located between the catalytic loop, the P+1 loop (activation segment), helix D, helix F, helix G and helix H.

#### 4.3.2.1 Cancer mutations in relation to sequence-conserved regions

We examined the distribution of the distances between mutated residues and conserved regions in the different protein kinases using two different definitions of sequence conservation. First, a simplistic approach was used based on Shannon's entropy [125] in the context of identity. Accordingly, 21 bins were characterized that each reflected an amino acid, with an extra bin for gaps. The positions in the multiple sequence alignment were labeled as conserved if their Shannon's entropy was less than 0.20. Additionally, to avoid unreliable results, alignment positions with more than 75% gaps were automatically discarded. Thus, a total of 20 fulfilled the requirements to be considered conserved and 6 of them were driver mutations. Figure 4.10 shows that drivers tended to locate closer to conserved regions than passenger mutations, which was also reflected in the $Xd$ values: $\Delta Xd$ = -1.23, Table 4.2.

We also considered an alternative measure of sequence conservation, AL2CO [126]. An advantage of this program over the earlier method is that AL2CO weights the sequences in order to correct for the unequal distances between the sequences in the alignment, and it uses a scoring

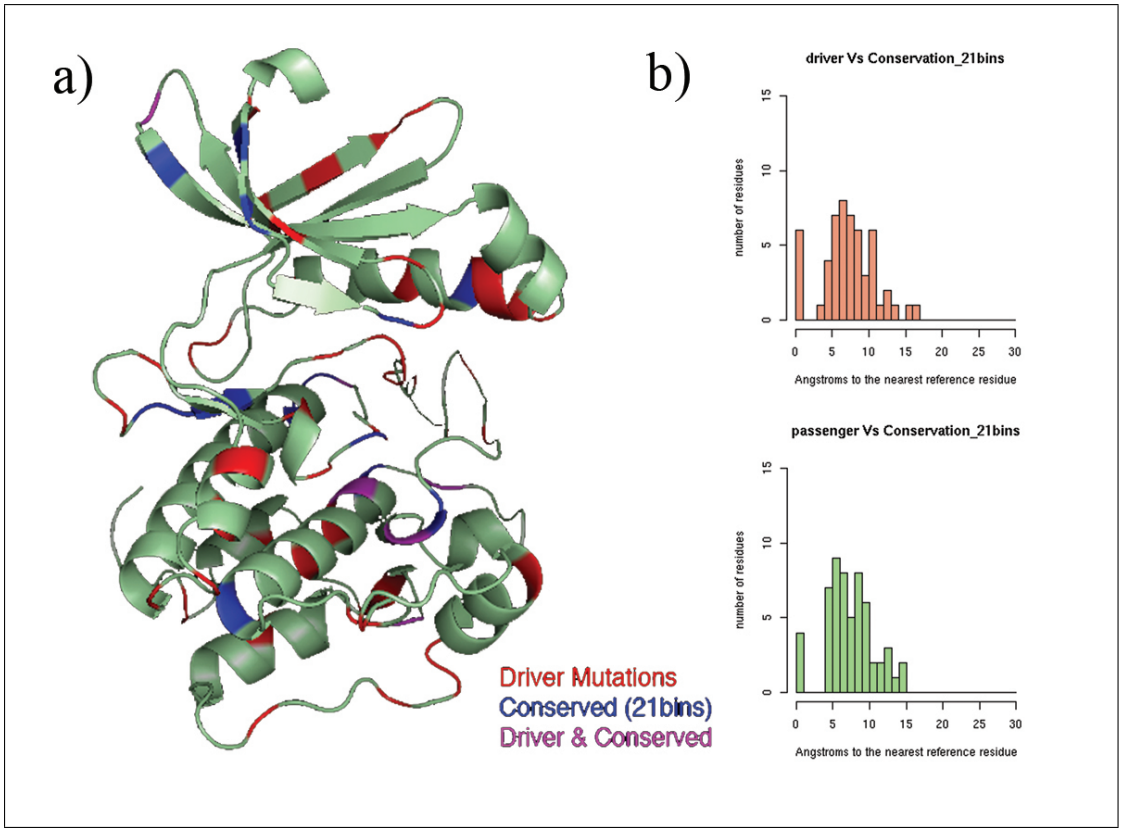



**Figure 4.10:** Driver and passenger mutations at the conserved sequence regions calculated in terms of Shannon's entropy.
Panel A: Driver mutations (red) and conserved positions (blue) mapped onto the consensus structural model. Positions with both a driver mutation and a conserved residue are displayed in magenta.
Panel B: The histograms show the distribution of distances between conserved positions and driver (red) or passenger (green) mutations.

matrix to consider the most common positions occupied by residues with similar physicochemical properties. As residues with a normalized conservation index threshold of 70% were deemed to be conserved, there were in total 14 conserved residues, of which 4 were driver mutations (Figure 4.11, panel A displays the position of these conserved residues and the driver mutations, with driver mutations tending to locate close to the conserved positions). Moreover, we observed that drivers were nearer to the conserved positions than passenger mutations ($\Delta Xd = -0.88$, Table 4.2 and Figure 4.11, panel B).

#### 4.3.2.2 Cancer mutations and solvent accessibility

We analyzed the distribution of the mutated positions with respect to 'buried' residues, which were defined as those with a relative solvent accessibility score of less than 16% when measured with Naccess (*Hubbard, unpublished*). Out of the 99 buried residues in the protein kinase domain, 16 and 17 residues coincided with driver and passenger mutations, respectively. Moreover, we found no differences in the weighted distribution of the distances between the two kinds of mutations ($\Delta Xd = -0.01$, Table 4.2).

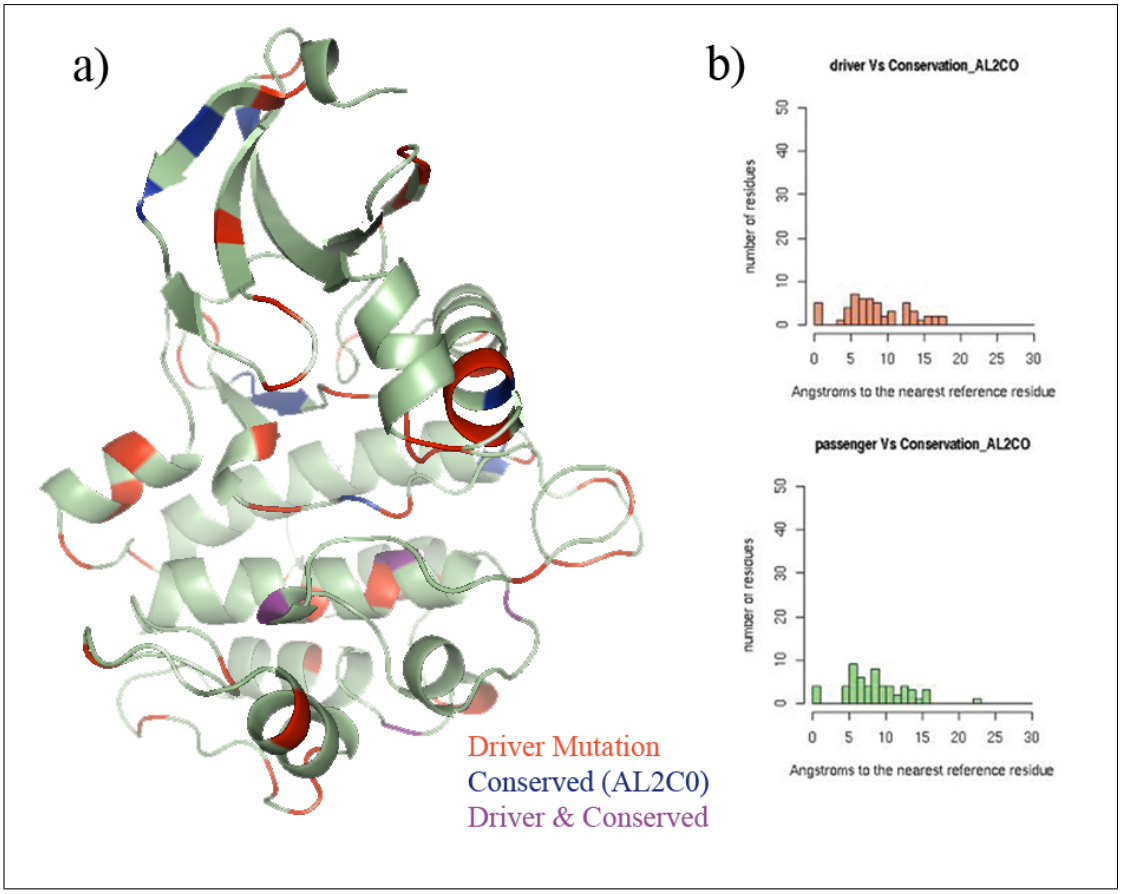



**Figure 4.11:** Driver and passenger mutations at the conserved positions calculated with AL2CO [126]. Panel A: Driver mutations (red) and conserved positions (blue) mapped onto the consensus model structure. Positions containing both a driver mutation and a conserved residue are displayed in magenta. Panel B: The histograms show the distribution of distances between the conserved positions and the driver (red) or passenger (green) mutations.

#### 4.3.2.3 Cancer mutations and the ATP-binding site

The functionally active region of protein kinases was defined in the Introduction. However, it is worth remembering that the ATP-binding pocket has three main parts:

1. A region of hydrophobic residues clustered around the adenosine of ATP;
2. An area around the $\gamma$-phosphate of ATP and the divalent cation (the catalytic site) that is primarily enclosed by charged residues;
3. A region in the large lobe composed of both hydrophobic and polar residues below the ATP that stabilizes this region and that may mediate substrate interactions.

There are 32 residues in FireDB [127] that describe the consensus ATP-binding pocket of the protein kinase superfamily (Figure 4.12), while others have limited the catalytic region of the kinases to a subset of these residues [46], considering only the five highly conserved residues that are responsible for ATP positioning and for stabilizing the active conformation in the catalytic mechanism:

1. Lys-74, which interacts with the alpha- and beta-phosphates of ATP and stabilizes it.
2. Glu-96, which forms a salt bridge with Lys-74 and increases the stability of this network.
3. Asp-171, which serves as the catalytic base, initiating phosphotransfer by deprotonating the acceptor serine, threonine or tyrosine.
4. Asn-176, which interacts with a secondary divalent cation, thereby positioning the $\gamma$-phosphate of ATP.
5. Asp-190, which chelates the primary divalent cation and indirectly positions ATP.

Driver mutations appeared near the ATP-binding pocket (Figure 4.12, panel A) and indeed, 8 out of the 32 residues in this pocket correspond to positions with a driver mutation. When the distance distribution of point mutations was examined (Figure 4.12, panels B and C), driver mutations were closer to the positions forming the ATP-binding pocket than passenger mutations, whether this pocket was defined by either FireDB or Knight and coworkers [46]. Nevertheless, this trend was stronger for the residues in direct contact with the substrate identified by FireDB, as supported by the differences in the harmonic deviations $\Delta Xd_{FireDB} = -0.95$ and $\Delta Xd_{Knight} = -0.55$ (Table 4.2) for the ATP-binding site.

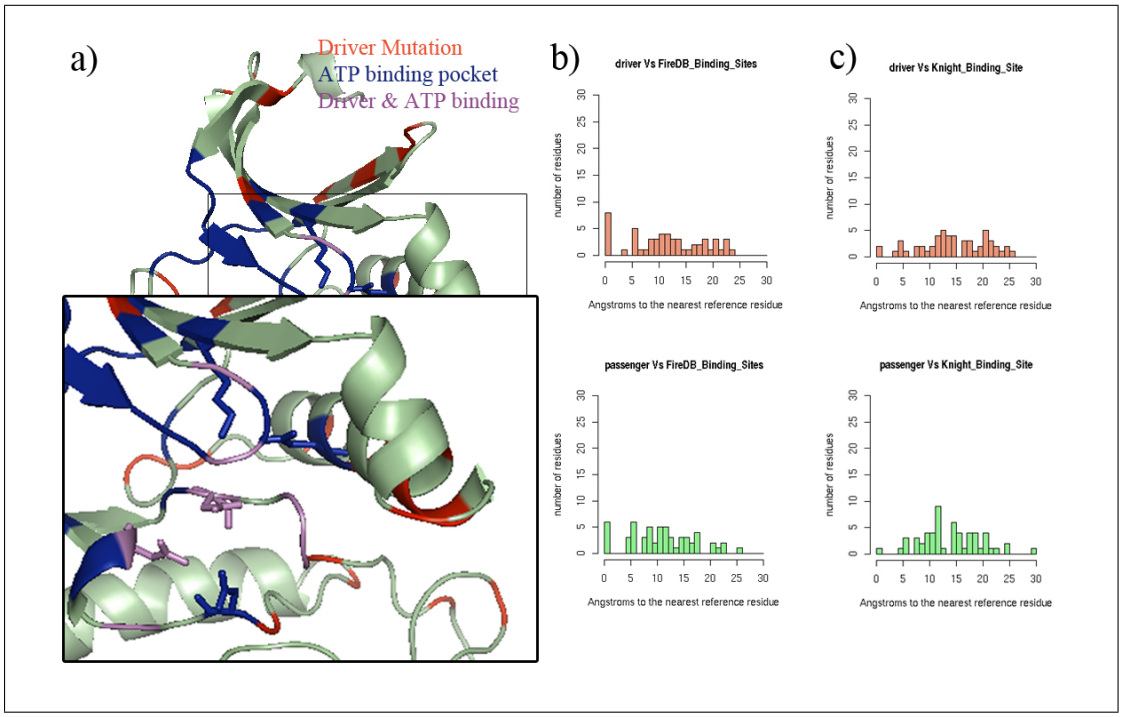



**Figure 4.12:** Driver and passenger mutations in the ATP-binding pocket.
Panel A: Driver mutations (red) mapped onto the ATP-binding pocket (blue). Positions that have a driver mutation and are part of the catalytic site are displayed in magenta. We have used the set of binding residues extracted from FireDB as an operational definition of the consensus ATP binding site of the kinases. This definition includes the five highly conserved residues mentioned by Knight [46]: K74, E96, E171, N176 and E190 (represented by sticks).
Panel B: The histograms show the distribution of distances between residues forming the ATP-binding pocket, according to FireDB [127], and driver (red) or passenger (green) mutations.
Panel C: The histograms show the distribution of distances between residues forming the ATP-binding pocket, according to Knight and coworkers [46], and driver (red) or passenger (green) mutations.

#### 4.3.2.4 Cancer mutations and tree-determinants

We define tree-determinants as residues putatively involved in binding specificity. They are often used as a proxy for functionally important regions in protein kinases, particularly those related to the specific functions of each of the groups, and therefore, they are also referred to as specificity-determining positions (SDPs) [129, 131, 157]. Residues differentially conserved in the various groups of protein kinases were identified for each of the 8 groups KinBase uses to categorize the human kinome [40] using the S3Det method [131]. We identified 35 residues as sufficient to differentiate between the groups, i.e., residues that are conserved within a specific kinase group but different from those of the rest of the groups. The most statistically significant tree-determinants were distributed among the kinase groups as follows: 4 in the AGC, 4 in CK1, 8 in CMGC and 9 in STE.

Tree-determinants were mapped onto the representative structure (Figure 4.13, panel A) and while a representative set of tree-determinants clustered near the ATP-binding pocket, others were found around the bundle of helices that form the C-lobe. This region is known to bind substrates, interact with other protein partners and participate in intermolecular signaling [158]. The distribution of point mutations with respect to the positions of the tree-determinants was determined (Figure 4.13, panel B) and the corresponding $Xd$ values were compared ($\Delta Xd = -1.55$, Table 4.2). Drivers were closer than passenger mutations to positions that are important in determining kinase group sub-specificity.

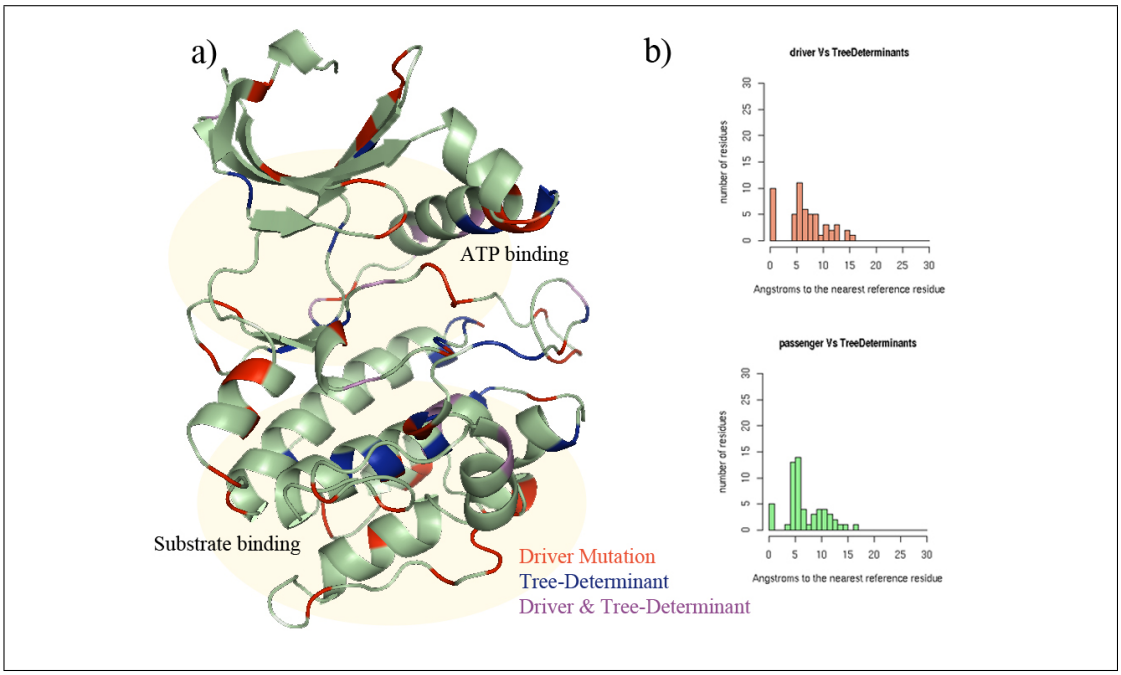



**Figure 4.13:** Driver and passenger mutations in specificity regions.
Panel A: Driver mutations (red) and tree-determinants (blue) mapped onto the consensus structural model. Positions with both a driver mutation and a tree-determinant are displayed in magenta. While a representative set of the Tree-Determinants clustered near the ATP binding pocket of the protein, others could be found around the bundle of helices that conform the C-lobe. This region is known to bind substrates, interact with other protein partners and to participate in intermolecular signaling events.
Panel B: The histograms show the distribution of distances between residues determining the kinase group sub-specificity and driver (red) or passenger (green) mutations.

### 4.3.3 Distribution of pathogenic germline mutations

Following a similar approach to the one previously reported for the driver and passenger somatic mutations, we mapped different types of germline mutations onto a representative structural model of the protein kinase superfamily and analyzed their distribution relative to evolutionary conserved positions and known functional regions. The two datasets used to compare pathogenic and neutral mutations were downloaded from SAAPdb. The pathogenic dataset corresponded to germline mutations classified as 'pathogenic deviations' (PDs) and more likely to be associated with disease, while the neutral dataset consisted of 'neutral SNPs' that are widespread in the population and not likely to be pathogenic. Both datasets were limited to mutations within the protein kinase domain.

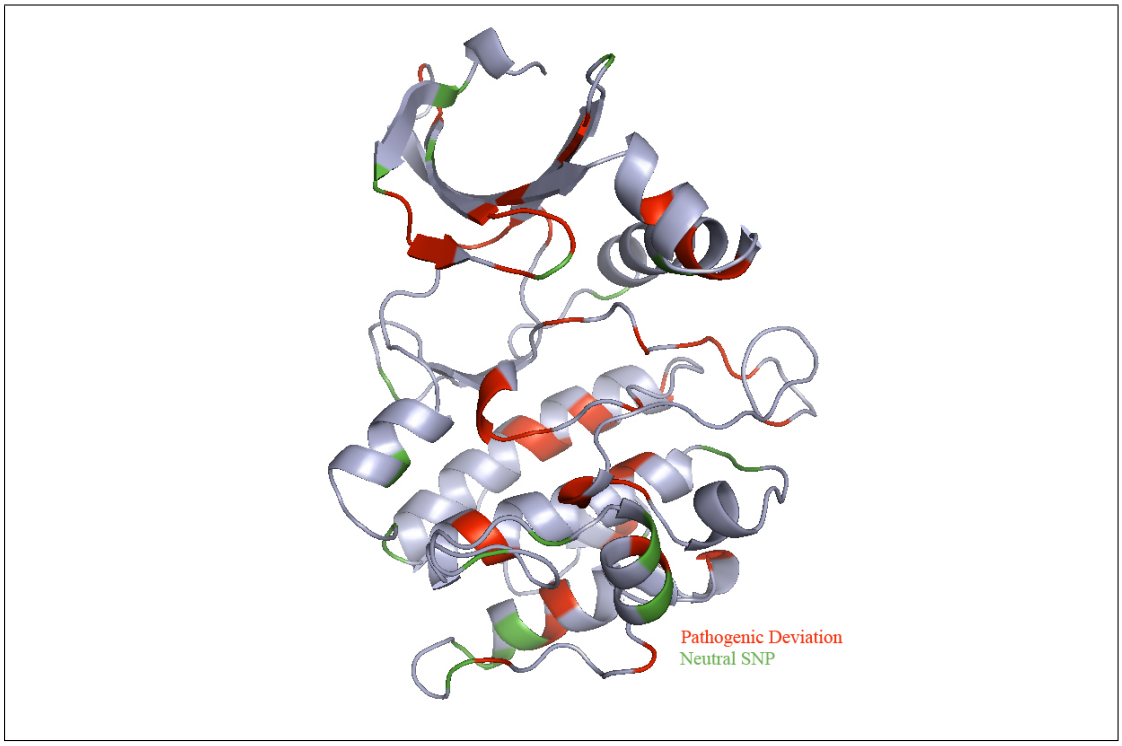



**Figure 4.14:** Our model structure of the human protein kinase domain based on MAP3K1. 47 positions with at least one pathogenic mutation (red), those that correspond to residues that are putatively associated with disease. By contrast, 27 positions had neutral SNPs (green).

We mapped 47 positions containing at least one PD and 27 positions with at least one SNP onto the consensus model described previously (Figure 4.14. We calculated the $Xd$ scores of the distribution of distances [133] to prioritize the differences in regions closer to those studied (for example binding sites) over the differences in the distribution of residues at positions far from the regions of interest. We undertook a similar approach to analyse somatic mutations. Again, to compare the two distributions, the $\Delta Xd$ was calculated as the difference of the $Xd_{SNP}$ and $Xd_{PD}$. Greater negative values indicated that the PDs and the positions under study co-localize, while greater positive values indicated the contrary. An arbitrary threshold of 0.5 was chosen as sufficient to discern between the classes (the results of these analyses are shown in Table 4.3).

| | **Mean Dist. PDs (Å)** | **Mean Dist. SNPs (Å)** | $Xd_{PD}$ | $Xd_{SNP}$ | $\Delta Xd$ |
|---|---|---|---|---|---|
| Seq. conservation, Shannon | 7.17 | 6.71 | -0.01 | -0.14 | -0.13 |
| Seq. conservation, AL2CO | 8.49 | 10.49 | 1.07 | 0.55 | -0.52 |
| Accessibility | 2.94 | 3.63 | 1.84 | 1.01 | -0.83 |
| Catalytic site, FireDB [127] | 8.69 | 12.66 | 3.92 | -0.84 | -4.76 |
| Catalytic site, Knight [46] | 11.89 | 16.26 | 1.35 | -0.98 | -2.33 |
| Tree-determinants | 6.99 | 7.94 | 0.31 | -1.67 | -1.98 |

**Table 4.3:** Distribution of PDs and SNPs in regions that are evolutionary conserved, display structural conservation or retain functionality

#### 4.3.3.1 Germline mutations and sequence conservation

Sequence conservation was evaluated with AL2CO [126] and as in the protocol described earlier for driver and passenger somatic mutations, residues with a normalized conservation index threshold of 70% were labeled as conserved. This analysis revealed that 3 out of the 14 conserved residues coincided with PDs and two with neutral SNPs. Indeed, conserved positions tended to be surrounded by PDs even if there were not many exact coincidences of conserved residues and PDs (Figure 4.15, panel A). In the histograms showing the distribution of the distances between mutated and conserved positions for PDs and neutral polymorphisms (Figure 4.15, panel B), in general PDs in the protein kinase domain were closer to conserved residues than neutral polymorphisms, as further supported by the $\Delta Xd$ of -0.52 (Table 4.3).

A similar trend was observed when analyzing Shannon's entropy. Positions in a multiple sequence alignment were deemed to be conserved if their Shannon's entropy was less than 0.20, and to ensure the results were more reliable, the alignment positions with more than 75% gaps were discarded. Under these constraints, 20 residues were identified as conserved and 4 of them coincided with the position of a PD, including the three residues observed with AL2CO and a glycine in the glycine-rich loop. In addition, the same two neutral SNPs were highly conserved (Figure 4.16, panel A). In the histograms of the distribution of the distances between mutated and conserved positions (Figure 4.16, panel B), the PDs in the protein kinase domain tended to be slightly closer to the conserved residues than neutral polymorphisms. The $\Delta Xd$ was small at only -0.13 (Table 4.3), corroborating that this trend was not very strong.

#### 4.3.3.2 Germline mutations and accessibility to the solvent

We identified 99 residues as inaccessible to the solvent under the same constraints defined for the driver and passenger somatic mutations, i.e., relative residue solvent accessibility score of less than 16% measured with Naccess (*Hubbard, unpublished*). Out of these 99 buried residues, 20 were PDs (Figure 4.17, panel A) and 12 were neutral SNPs. The distribution of distances (Figure 4.17, panel B) clearly demonstrates PDs being closer to buried residues in the protein core. Analyzing the distance between mutated positions and solvent-inaccessible residues revealed that PDs were closer to buried residues than neutral polymorphisms, supported by an $Xd$ difference of -0.83 (Table 4.3).

#### 4.3.3.3 Germline mutations and catalytic sites

The ATP-binding pocket of the kinase superfamily was defined as the set of residues extracted from FireDB [127], including the 32 residues (Figure 4.18, panel A) that directly contact ATP in the binding pocket. This definition includes the five highly conserved residues that play a critical role in positioning ATP and stabilizing the active conformation in the catalytic mechanism [46].

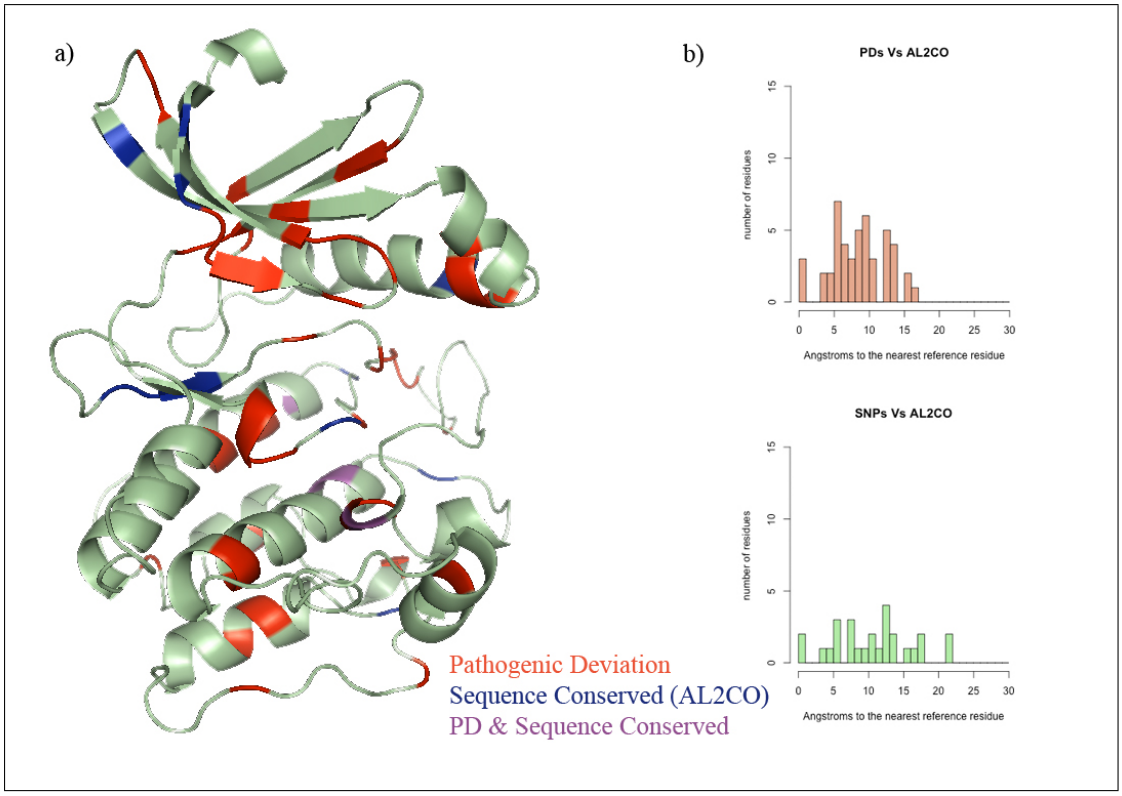



**Figure 4.15:** Pathogenic deviations (PD) and neutral polymorphisms (SNP) at conserved positions in the protein kinase domain calculated with AL2CO [126].
Panel A: Pathogenic deviations (red) and conserved positions (blue) mapped onto the consensus model structure. Positions with both a pathogenic mutation and a conserved residue are displayed in magenta. Panel B: The histograms show the distribution of distances between conserved positions and pathogenic deviations (red) or neutral SNPs (green).

Regardless of the definition used (FireDB or Knight), PDs tended to locate in the catalytic region or at least around it. Indeed, 13 out of the 32 residues were annotated as PDs, whereas only 3 were neutral SNPs. Moreover, two of the five residues thought to be essential for the correct functioning of the ATP-binding pocket [46] were annotated as PDs. Both the distance distribution histograms (Figure 4.18, panels B and C) and the $\Delta Xd$ results ($\Delta Xd_{FireDB} = -4.76$ and $\Delta Xd_{Knight} = -2.33$, Table 4.3) support this trend.

#### 4.3.3.4 Germline mutations and regions of functional sub-specificity

As for driver and passenger somatic mutations, we also investigated the relationship between tree-determinants and germline mutations. Tree-determinants are a proxy for functionally important regions in protein kinases, particularly those related with the specific functions of each one of the eight groups in which KinBase classifies the human kinome [40]. Consequently, they are often referred as subspecificity determining positions (SDPs) [129, 131, 157]. Our recent automation of the sequence-space approach, S3Ddet [131], identified 32 unique positions as discriminative between kinase specificity groups. A number of these residues clustered close to the ATP-binding pocket, while others were concentrated around the bundle of helices in the C-lobe, which is known to bind substrates, interact with other protein partners and participate in intermolecular signaling [158].

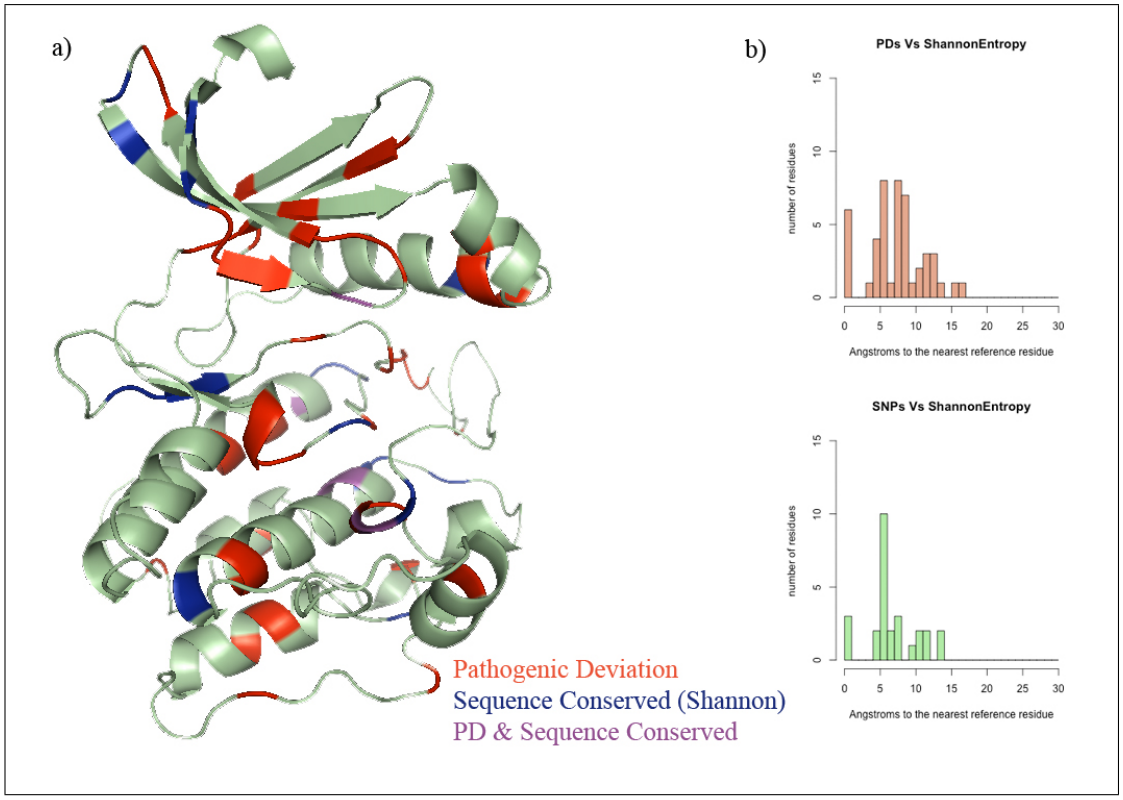



**Figure 4.16:** Pathogenic deviations (PD) and neutral polymorphisms (SNP) at conserved positions in the protein kinase domain calculated in terms of Shannon's entropy [159].
Panel A: Pathogenic deviations (red) and conserved positions (blue) mapped onto the consensus model structure. Positions with both a pathogenic mutation and a conserved residue are displayed in magenta.
Panel B: The histograms show the distribution of distances between conserved positions and pathogenic deviations (red) or neutral SNPs (green).

Pathogenic deviations located closer to tree-determinant positions than neutral SNPs. PDs, if not exactly in positions that were annotated as tree-determinants, clustered around specificity residues. In fact, out of the 32 tree-determinants mapped onto the consensus model, five were annotated as pathogenic and only two as neutral. This is especially relevant for residues around the ATP-binding pocket but it was also appreciable in the other function specific tree-determinants (Figure 4.19, panel A depicts the localization of pathogenic deviations and kinase specificity-determining residues). There was also a clear difference in $Xd$ values ($\Delta Xd = -1.98$, Table 4.3) and the histograms showing the distribution of distances (Figure 4.19, panel B), indicating that significant differences between PDs and SNPs exist regarding their proximity to positions characterized as important for kinase function and sub-specificity.

### 4.3.4 Assessing the possible functional role of relevant kinase mutations by their sequence-structure characteristics

In the previous section, we mapped all the knowledge accumulated for the protein kinase domain onto a consensus model under the assumption that regions important for kinase structure and function can be used as a reference to interpret the mutations in this superfamily. The accumulation of information clearly increases the significance of the results provided and makes

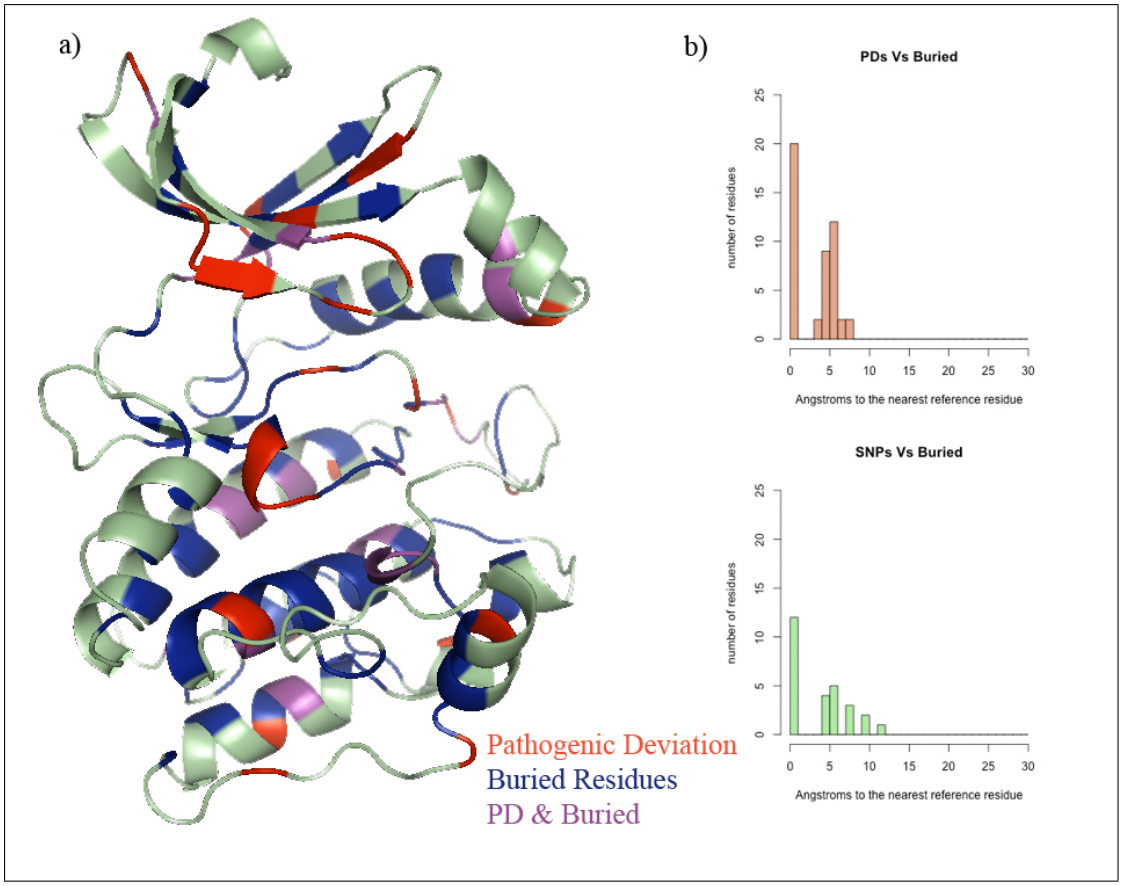



**Figure 4.17:** Pathogenic deviations (PD) and neutral polymorphisms (SNP) in regions of the protein that are solvent-inaccessible and form the core of the protein.
Panel A: Pathogenic deviations (red) and buried positions (blue) mapped onto the consensus structural model. Positions with both a pathogenic mutation and a buried residue are displayed in magenta.
Panel B: The histograms show the distribution of distances between solvent-inaccessible positions and pathogenic deviations (red) or neutral SNPs (green).

the distribution of polymorphisms more reliable and accurate. Once the relevance of the analyzed regions in the structure was understood, our approach also provided an insight into the specific biomedical implications of individual kinase mutations.

#### 4.3.4.1 Diabetes, acanthosis nigricans and mutations in the insulin receptor

Mutations in the human insulin receptor gene (INSR) have been associated with disease and defects in the INSR have been linked to patients suffering from insulin-resistant diabetes mellitus associated with acanthosis nigricans type A (IRAN type A, MIM: 610549). IRAN type A is characterized by severe insulin resistance manifesting as marked hyperinsulinemia, no response to exogenous insulin, ovarian hyperandrogenism in adolescent female patients and the skin lesion acanthosis nigricans. The relationship between this disease and the mutations in kinases is reflected in OMIM and the literature [150]. Recent studies further characterized this relationship between insulin resistance and disease [160], while acanthosis nigricans has been linked to mutations in other kinases, such as fibroblast growth factor receptors II and III [161–163].

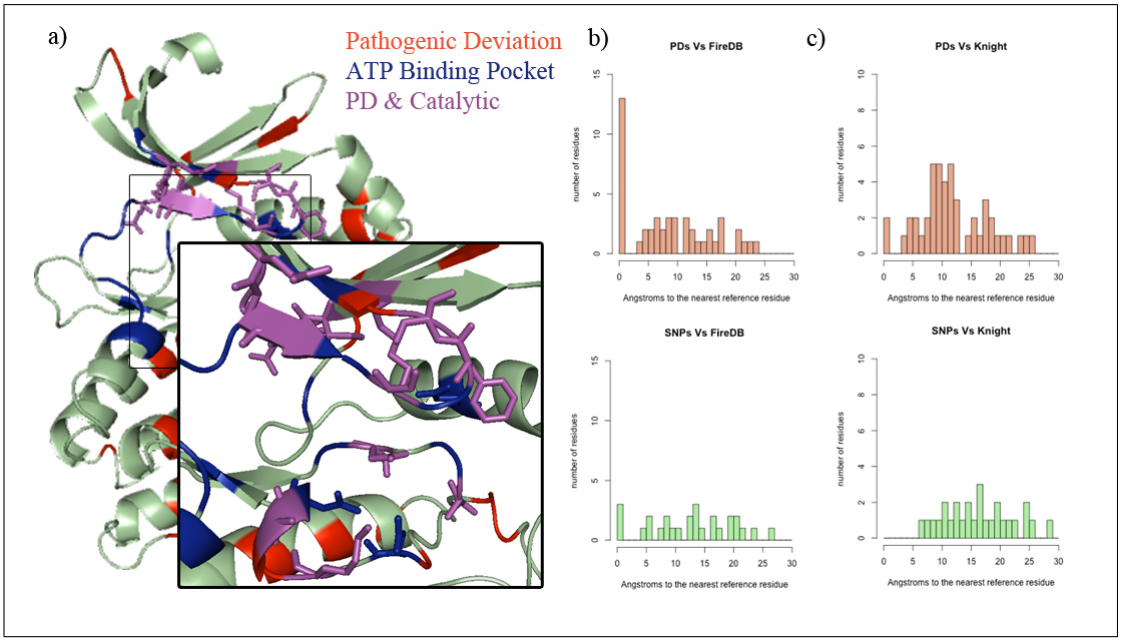



**Figure 4.18:** Pathogenic deviations (PD) and neutral polymorphisms (SNP) in the ATP-binding pocket.
Panel A: Pathogenic deviations (red) and conserved positions (blue) mapped onto the consensus model structure. Positions with both a pathogenic mutation and a buried residue are displayed in magenta. We used the set of residues extracted from FireDB as an operational definition of the binding site. This definition includes the five highly conserved residues mentioned in Knight et al. (2007) [46]: K74, E96, E171, N176 and E190 (represented by sticks).
Panel B: The histograms show the distribution of distances between residues forming the ATP-binding pocket, according to FireDB [127], and pathogenic deviations (red) or neutral SNPs (green).
Panel C: The histograms show the distribution of distances between residues forming the ATP-binding pocket, according to Knight and co-workers [46], and pathogenic deviations (red) or neutral SNPs (green).

We were interested in identifying disease-prone mutations based on the accumulation of pathogenic factors in the position of the mutations. For instance, we identified A1161T, a mutation that introduces an alanine-threonine shift in the INSR. This mutation is at a catalytic residue according to FireDB [127] and it is important for family specificity [131]. Perturbing the ATP-binding pocket might impair phosphorylation and reduce enzymatic activity, thereby producing an aberrant phenotype. To corroborate this prediction, we found that this mutation has been defined in SAAPdb [74] as a pathogenic deviation.

#### 4.3.4.2 The carcinogenic role of mutations in B-raf

Another interesting example is the mutation D594G in B-raf proto-oncogene. The RAF gene family has three members (ARAF1, BRAF and RAF1), each encoding a serine/threonine kinase that is regulated by binding to RAS, as component of the RAS-RAF-MEK-ERK-MAP kinase pathway that plays a critical role in cell proliferation. B-raf is frequently activated in cancer cells, especially in non-Hodgkin lymphoma (NHL) and mutated B-raf proteins have been shown to have greater kinase activity [148], indicating that the RAS-RAF kinase pathway is probably regulated by somatic mutations of B-raf in some cancers.

Here, we identified D594G as a pathogenic mutation that introduces an Asp/Gly change in

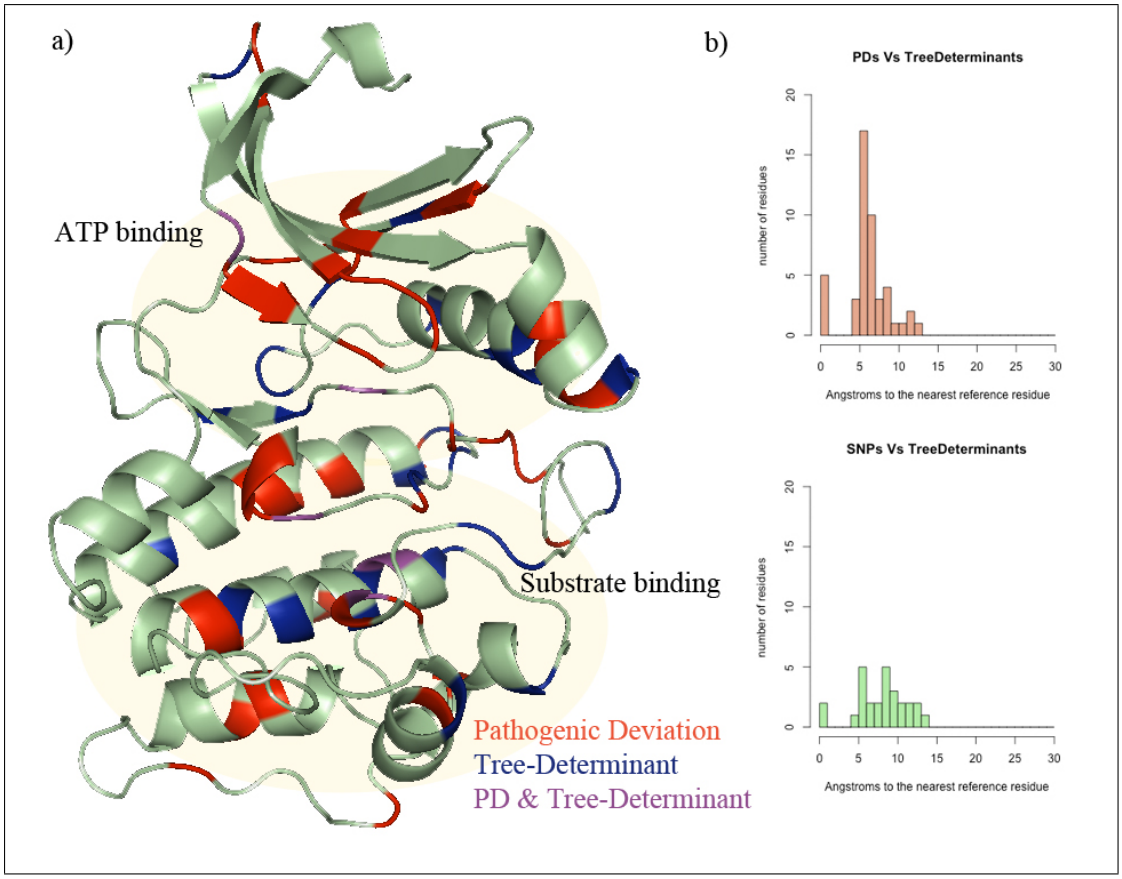



**Figure 4.19:** Pathogenic deviations (PD) and neutral polymorphisms (SNP) in regions of kinase-group sub-specificity.
Panel A: Pathogenic deviations (red) and tree-determinants (blue) mapped onto the consensus model structure. Positions with both a pathogenic mutation and a specificity-determining residue are displayed in magenta.
Panel B: The histograms show the distribution of distances between tree-determinants and pathogenic deviations (red) or neutral SNPs (green).

the activation loop of the kinase family. Our analysis highlighted the importance of this amino acid in ATP binding, as described by FireDB [127], and as a residue conferring family sub-specificity [131]. From these descriptions, we derived the hypothetical role of this mutation in the development of NHL as described above. To confirm this hypothesis, SAAPdb [74] and our text-mining pipeline [150] indicated that this mutation is pathogenic and involved in the disruption of the binding site, interacting surface, quaternary structure and the essential scaffolding of hydrogen bonds.

## 4.4 Prioritizing pathogenic mutations in the protein kinase superfamily

Human protein kinases fulfill a wide variety of physiological functions and while most mutations described in this family are tolerated without significantly disrupting the corresponding structures or functions, some have been linked to a variety of human diseases, including cancer.

We have already discussed the preferential distribution of germline pathogenic deviations [155] and driver somatic mutations [154] to regions of functional and structural importance. Here, we present the basis for the development of a computational method to predict the impact of mutations on the function of protein kinases based on these features.

We explored the significance of disease-associated mutations in terms of sequence-derived characteristics at different levels:

1. *At the gene level:* membership to a KinBase group and Gene Ontology terms.

2. *At the domain level:* the occurrence of the mutation inside a PFAM domain.

3. *At the residue level:* several properties including amino acid types, functional annotations from SwissProt and FireDB, and specificity-determining positions (SDPs).

Accordingly, we analyzed the independent significance of these properties and their combination with a support vector machine (SVM).

### 4.4.1 Construction of the disease and neutral datasets

The method was trained and evaluated using a dataset derived from UniProt [72], which has been benchmarked for a number of classifiers with satisfactory results [136]. After the filtering pipeline described in the Methods, 865 mutations in 65 human kinases formed the 'disease dataset', whereas the 'neutral dataset' consisted of 2,627 mutations in 447 human kinases.

### 4.4.2 Optimization of the prediction method

To classify the mutations in the human kinome as disease-associated or neutral according to the sequence features of the mutations, we used a Support Vector Machine (SVM). This type of approach has previously been widely used to automatically prioritize disease-associated mutations [102,107,108,140,164] and it has been demonstrated to outperform other approaches such as Bayesian classifiers and neural networks [107].

Our implementation of the SVM relied on a radial basis function kernel. Two parameters are crucial for the performance of the classifier, the soft-margin penalty ($C$) and the radius ($\gamma$): $C$ represents the amount of errors allowed during the training and evaluation steps, while $\gamma$ represents the width of the SVM radial function. These parameters can be optimized to improve the predictions and consequently, the parameters were exhaustively evaluated for values ranging between $0 \leq C \leq 8$ and $10^{-4} \leq \gamma \leq 10^{-2}$ (Figure 4.20). To decide which parameters predict with the best performance, the average f-score across the entire set of k-folds was chosen as a scoring function for optimization (the whole procedure is described in the Methods). The optimal values used during the analyses were $C = 8$ and $\gamma = 6 \cdot 10^{-4}$

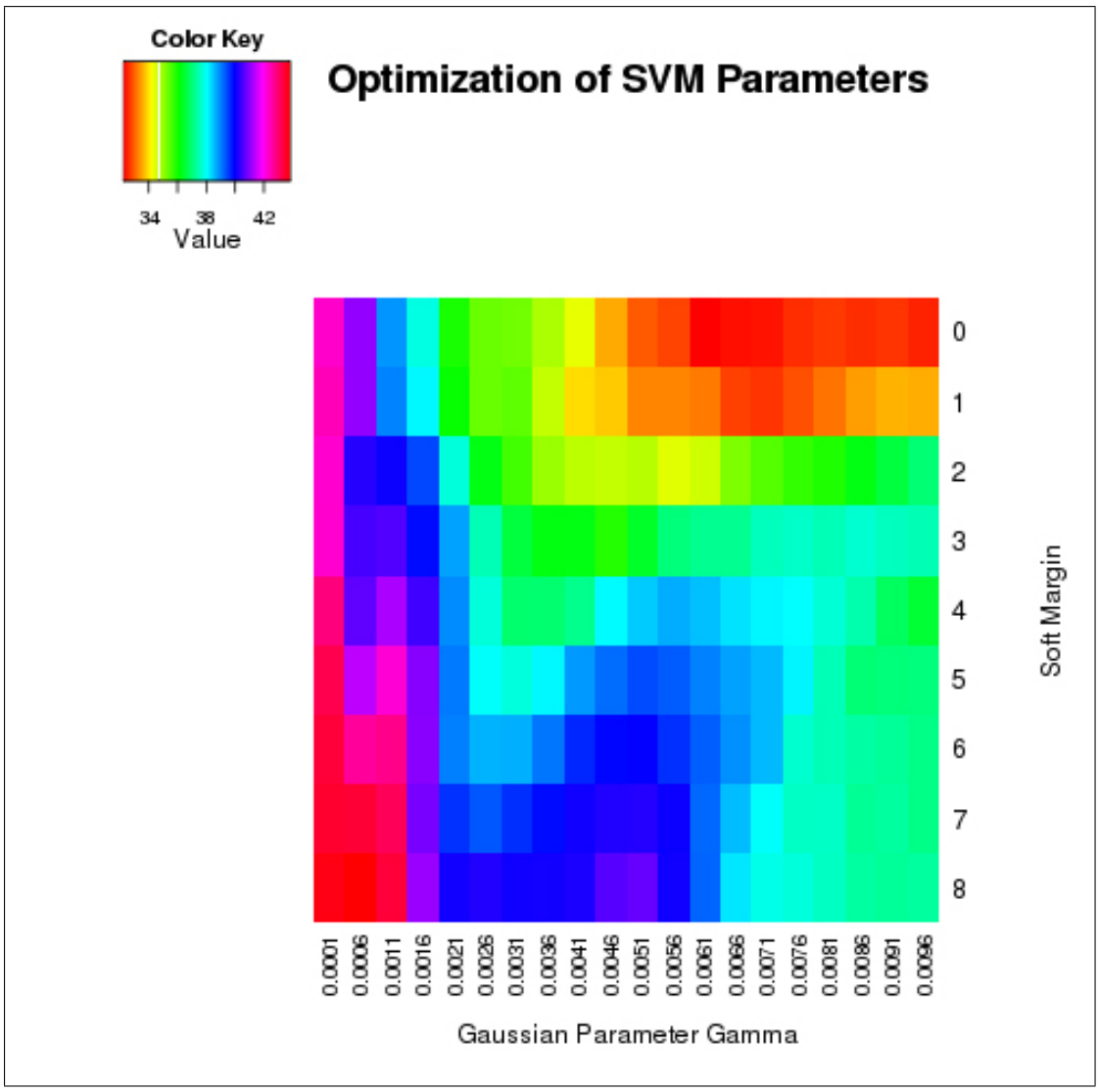



**Figure 4.20:** Grid optimization of the predictive power of the classifier. We exhaustively tested the two most critical parameters of the SVM's radial basis kernel: soft-margin ($C$) and radius ($\gamma$). The average f-score across the entire set of k-folds was chosen as a scoring function for the optimization. The optimal values used for the analyses were $C = 8$ and $\gamma = 6 \cdot 10^{-4}$ when all groups in the kinase superfamily were considered.

### 4.4.3 Evaluation of the performance of the classifier

We avoided over-fitting the classifier by applying a 10-fold cross-validation approach where 8 random sets were used during the training step, one for the validation phase and one for the evaluation. This process was repeated 10 times to allow all the sets to be used during the evaluation. Although the optimization of the kernel relies on the f-score, the performance of the classifier is assessed by several additional measures, such as accuracy, precision, recall and the Matthew's correlation coefficient (MCC).

On average, the classifier predicted the pathogenicity of kinase mutations very accurately. Different threshold values could modulate the output of the classifier, as summarized

in Table 4.4, and selecting an appropriate threshold is a critical step in developing a classifier. Relaxed thresholds, such as -0.75, enable the detection of more disease-associated mutations (increased recall), albeit at the cost of a larger number of false positives (reduced precision). Conversely, higher thresholds of conservative classifiers, such as -0.5, reduce the frequency of a mutation being classified as pathogenic, consequently predicting a smaller set of more reliable disease-associated mutations. For the sake of accuracy, we analyzed our results according to the conservative threshold of -0.5, whereby which the classifier predicted 83.29% of the mutations correctly. Regarding the pathogenic dataset, 75.17% of the observed mutations were recovered on average across all k-folds with a precision of 60.03%. The average MCC was 0.58.

| SVM threshold | Accuracy (%) | Precision (%) | Recall (%) | MCC |
|---|---|---|---|---|
| -1.00 | 74.33 | 46.52 | 89.14 | 0.65 |
| -0.75 | 80.91 | 56.12 | 79.84 | 0.62 |
| **-0.5** | **83.29** | **60.03** | **75.17** | **0.58** |
| -0.25 | 80.55 | 59.98 | 66.14 | 0.51 |
| 0.00 | 82.26 | 58.56 | 47.19 | 0.38 |

**Table 4.4:** Performance of the classifier depending on the SVM classification thresholds applied using all kinase groups.

### 4.4.4 Evaluation of the results in a data populated subset

There is a clear bias in the distribution of mutations across the kinase groups, since a small number of groups contain most of the mutations (Figure 4.21). Consequently, we evaluated the dependence of the results on the amount of disease-associated mutations available.

When the different groups in which the protein kinase superfamily is divided were superimposed (Figure 4.21), we observed differences in the number of mutations that populated the groups. This observation is consistent with the phylogenetic distribution of the literature-extracted mutations we discussed earlier (Figure 4.6). A small number of these groups contain most of the mutations, while others lack or contain very few disease-associated mutations (Table 4.5). For the mutations in these less populated groups, only group membership suffices to consider them as neutral and this neutrality is likely an artifact due to the lack of experiments assessing the pathogenicity of the mutations.

A second dataset was generated with only the highly populated groups: TK, TKL, Atypical_PI3-PI4, CAMK, RGC, CMGC, AGC and Atypical_ADCK. Under this constraint the 'disease dataset' consisted of 814 mutations in 54 human kinases, while the neutral dataset contained 1,775 in 297 proteins.

When only the groups sufficiently populated with disease-associated mutations were considered, on average we correctly predicted 76.81% of the remaining mutations across all k-folds, with the optimized values of $C = 8$ and $\gamma = 10^{-4}$ (Figure 4.22).

With respect to the pathogenic dataset, we recovered 73.26% of the disease-associated mutations with a precision of 64.68% (MCC: 0.53, Table 4.6), comparable to that obtained when all the mutations from all the kinase groups were considered, thereby confirming that the bias in the data does not significantly affect the results.

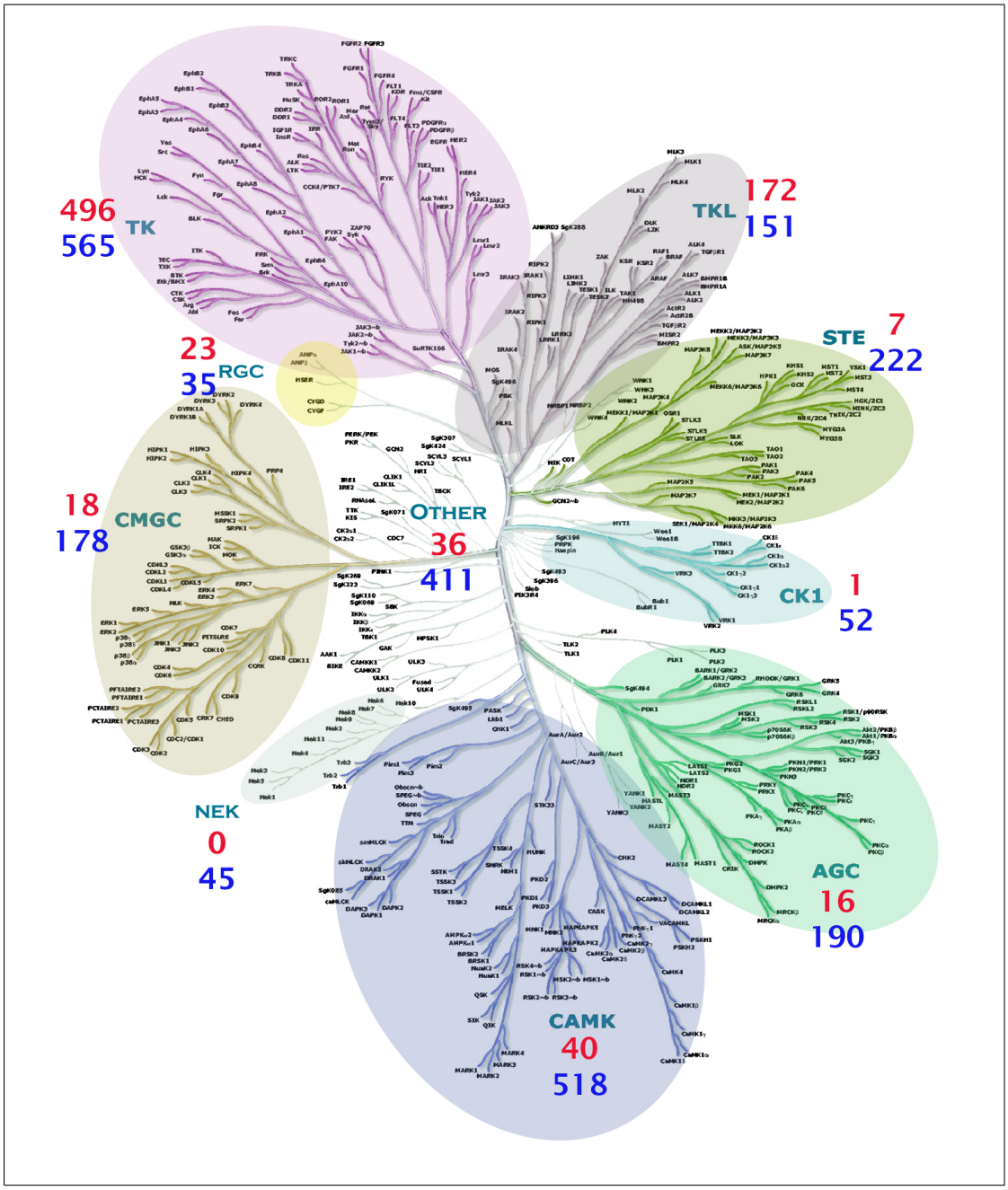



**Figure 4.21:** Mutations in each of the groups in which UniProt divides the protein kinase superfamily. The number of pathogenic (red) and neutral (blue) mutations is also represented. Differences in the number of mutations that populated the groups can be observed.

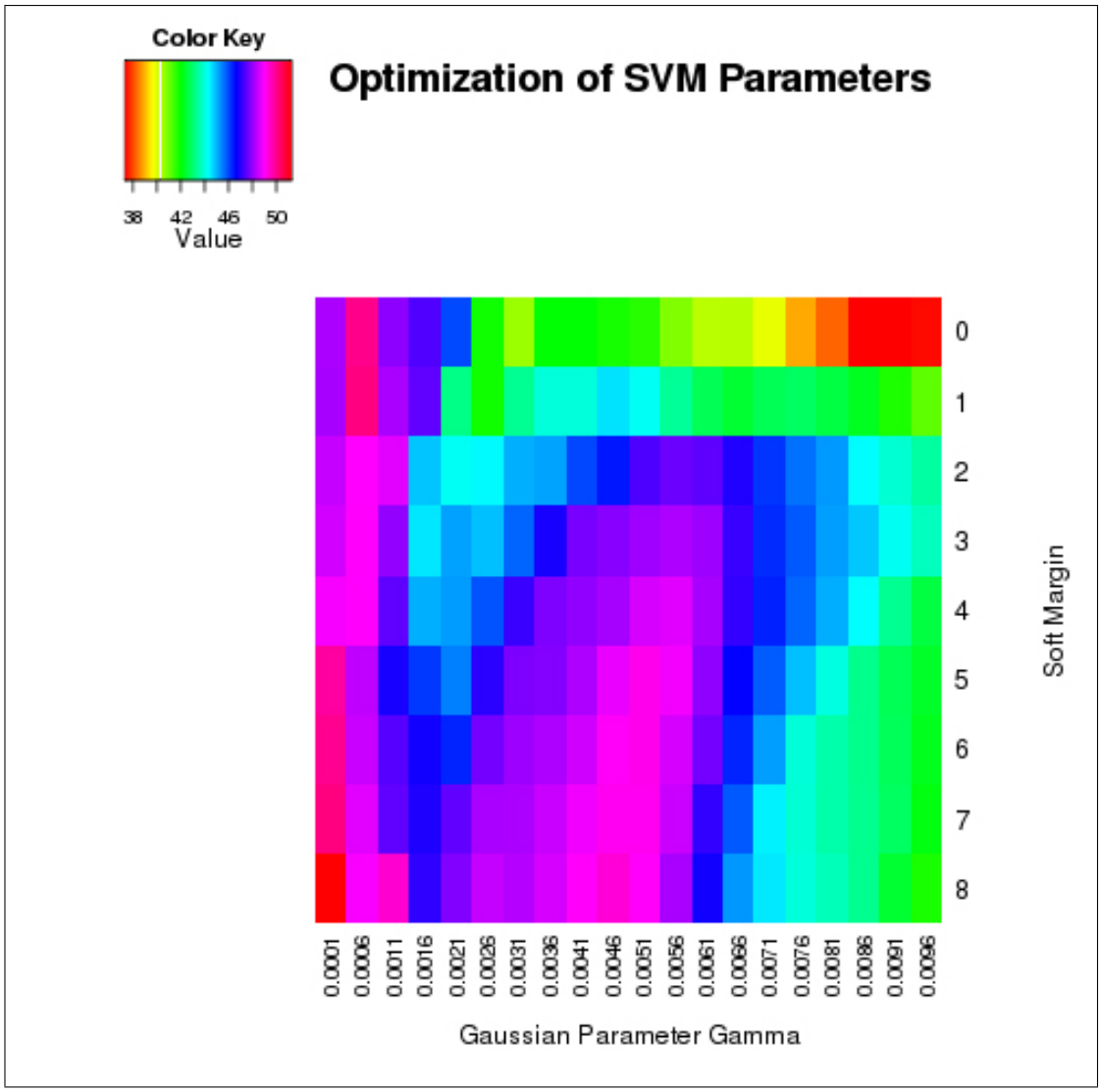



**Figure 4.22:** Grid optimization of the predictive power of the classifier when only the groups with a reasonable number of reported disease-associated mutations are considered. We exhaustively tested soft-margin ($C$) and $\gamma$. The average f-score across the entire set of k-folds was chosen as the scoring function for the optimization. The optimal values used during the analyses were $C = 8$ and $\gamma = 10^{-4}$.

| Group | Disease | Neutral | Total |
|---|---|---|---|
| TK † | 496 | 565 | 1061 |
| TKL † | 172 | 151 | 323 |
| Atypical_PI3-PI4 † | 49 | 138 | 187 |
| CAMK † | 40 | 518 | 558 |
| Other | 36 | 411 | 447 |
| RGC † | 23 | 35 | 58 |
| CMGC † | 18 | 178 | 196 |
| AGC † | 16 | 190 | 206 |
| STE | 7 | 222 | 229 |
| Atypical_ADCK | 6 | 14 | 20 |
| Atypical_Alpha-type | 1 | 88 | 89 |
| CK1 | 1 | 52 | 53 |
| NEK | 0 | 45 | 45 |
| Atypical_RIO | 0 | 14 | 14 |
| Atypical_PDK-BCKDK | 0 | 5 | 5 |
| Atypical_FAST | 0 | 1 | 1 |

**Table 4.5:** Number of mutations in each of the groups in which UniProt divides the protein kinase superfamily. The groups enriched in disease-associated mutations are highlighted by †.

| SVM threshold | Accuracy (%) | Precision (%) | Recall (%) | MCC |
|---|---|---|---|---|
| -1.000 | 71.50 | 51.43 | 88.93 | 0.61 |
| -0.750 | 77.06 | 61.72 | 81.55 | 0.60 |
| **-0.500** | **76.81** | **64.68** | **73.26** | **0.53** |
| -0.250 | 79.44 | 68.15 | 69.93 | 0.50 |
| 0.000 | 71.57 | 60.66 | 56.32 | 0.37 |

**Table 4.6:** Performance of the classifier depending on the SVM classification thresholds applied when using groups highly populated in disease mutations only.

However, clear differences were observed when the groups were compared individually (Table 4.7). For the groups with a reasonable number of mutations, the performance of the classifier was considerably better than with the less populated groups. This was especially clear for the precision of the predictions, which was consistent with the fact that the use of a sufficient number of support vectors helps the classifier learn how to discern disease-associated mutations properly.

### 4.4.5 Analysis of the most relevant features for classification

To evaluate the contribution of each individual feature to the classification, the features were ranked according to the variation in the module of the weight vector of the trained SVM ($\| \omega \|^2$) when each feature was removed from the set of support vectors. The feature whose removal minimized the variation in module was considered to contribute the least to the hyperplane that separates the two classes of examples (pathogenic/neutral) with a maximum margin (this ranking is shown in Table 4.8). The ranking derived from the SVMs has been applied for variable selection in many classification problems [165, 166]. According to the SVM-based criteria, the top ranked features are only based on the subset of support vectors

| Group | Accuracy (%) | Precision (%) | Recall (%) | MCC |
|---|---|---|---|---|
| CMGC † | 91.5 | 87.5 | 8.64 | 0.08 |
| TKL † | 68.7 | 70.47 | 70.93 | 0.37 |
| TK † | 71.3 | 69.68 | 68.32 | 0.42 |
| RGC † | 58.24 | 47.92 | 61.35 | 0.19 |
| Atypical_PI3-PI4 † | 70.59 | 47.12 | 100 | 0.78 |
| STE | 96.85 | 43.75 | 11.11 | 0.11 |
| AGC † | 90.78 | 43.35 | 61.11 | 0.55 |
| Other | 88.86 | 41.64 | 95.37 | 0.89 |
| CK1 | 97.69 | 33.33 | 22.22 | 0.21 |
| Atypical_Alpha-type | 89.89 | 9.09 | 88.89 | 0.83 |
| CAMK † | 55.48 | 8.31 | 51.94 | 0.1 |
| Atypical_ADCK | 70 | 0 | 0 | 0 |
| NEK | 100 | 0 | 0 | 0 |
| Atypical_RIO | 100 | 0 | 0 | 0 |
| Atypical_PDK-BCKDK | 100 | 0 | 0 | 0 |
| Atypical_FAST | 100 | 0 | 0 | 0 |

**Table 4.7:** Performance of the classifier when the groups in which UniProt divides the protein kinase superfamily are considered individually. Groups enriched in disease-associated mutations are indicated by †.

that are 'borderline' cases.

Interestingly, Gene Ontology (GO) functional annotations were the most relevant feature for the classification, which corresponds to a classification at the gene level since all the mutations in a gene display the same score. This information is encoded as the sum of GO terms log-odds ratio, sumGOLOR, to be able to compare between the classes. This measure represents the proportion of disease-associated and neutral kinase genes that are annotated with a particular GO term, and it can be used to identify the GO terms characteristic of neutral or disease-prone genes.

If the individual terms from the biological process sub-ontology are analyzed, interesting trends can be observed. For example, the most pathogenic biological processes are enriched in terms associated with protein localization, cell proliferation and tissue development, all aspects related to disease and particularly cancers. Pathogenic and neutral genes are differentially enriched in terms from the molecular function sub-ontology. While neutral genes are associated with basic kinase activity functions, disease-associated genes are enriched in terms associated with hormone binding, co-factors and interaction partners.

The most representative GO terms for each of the classes are shown in Table 4.9 for neutral genes and Table 4.10 for disease-associated ones.

The next group of features in the order of relevance for the predictor is linked to the positions that confer specificity at the family level (i.e., the tree-determinants). The calculation of this score is based on our in-house implementation of the S3Det method [131]. However, the current implementation of the method did not provide a continuous measure of tree-determinant characteristics and thus, we implemented this additional possibility. The coincidence of a given residue with the alignment of the rest of the family members, and the differences regarding the sequences outside the subfamily, were measured with an f-score as described in the Methods. Three different scores were calculated: the f-score for the wild type

| Rank | Feature Description | Rank | Feature Description |
|---|---|---|---|
| 1 | sumGOLOR | 21 | pfam_PF07714 (PKinase_tyr) |
| 2 | TDs_fscore_diff | 22 | phosphoelm |
| 3 | TDs_fscore_wt | 23 | class_CK1 |
| 4 | Diff_KD_hydrophobicity | 24 | swannot_disulfid |
| 5 | TDs_fscore_mt | 25 | SIFTscore |
| 6 | aatypeL | 26 | aatypeE |
| 7 | pfam_any | 27 | aatypeW |
| 8 | SIFTscore_binned | 28 | aatypeF |
| 9 | aatypeA | 29 | pfam_PF00041 (Fibronectin) |
| 10 | class_TK | 30 | class_TKL |
| 11 | aatypeV | 31 | class_Atypical_Alpha-type |
| 12 | aatypeR | 32 | aatypeQ |
| 13 | aatypeS | 33 | firedb |
| 14 | aatypeK | 34 | aatypeD |
| 15 | swannot_any | 35 | class_Atypical_ADCK |
| 16 | aatypeH | 36 | aatypeI |
| 17 | aatypeN | 37 | pfam_PF00023 (Ank) |
| 18 | aatypeM | 38 | pfam_PF01403 (Sema) |
| 19 | class_CAMK | 39 | class_Atypical_PI3-PI4 |
| 20 | aatypeT | 40 | aatypeG |

**Table 4.8:** Ranking of the features according to their contribution to the classification. The features are ranked according to the variation in the module of the weight vector of the trained SVM ($\| \omega \|^2$) when each feature is removed from the set of support vectors [165, 166]. See the Methods section for a detailed description of these features.

amino acid; the f-score for the mutant residue; and the difference between these two scores as a measure of the relevance of the change introduced.

Following these two important features of the predictor is the Kyte-Doolittle hydrophobicity change, the presence of a PFAM domain (in particular the tyrosine kinase domain), the functional annotation of the residues in SwissProt [72] and PhosphoELM [142], and the evolutionary SIFT score [93] or the amino acid types involved in the change.

Interestingly, among the genome-wide features, some kinase-specific features also emerged as being relevant. For instance, to reinforce the important role of gene-level characterization, classifying kinases into the different groups in KinBase [40] was an important feature (particularly TK, CAMK, CK1 and TKL among the canonical protein kinases, and Alpha-type, ADCK or PI3-PI3 among the atypical ones), as observed previously [107].

| GO term | %dis | %neu | GOLOR | Description |
|---|---|---|---|---|
| GO:0019901 | 0 | 4.62 | -32.1 | (MF) protein kinase binding |
| GO:0046328 | 0 | 4.62 | -32.1 | (BP) regulation of JNK cascade |
| GO:0070302 | 0 | 4.62 | -32.1 | (BP) regulation of stress-activated protein kinase signaling cascade |
| GO:0005083 | 0 | 3.85 | -31.84 | (MF) small GTPase regulator activity |
| GO:0043506 | 0 | 3.85 | -31.84 | (BP) regulation of JUN kinase activity |
| GO:0008134 | 0 | 3.59 | -31.74 | (MF) transcription factor binding |
| GO:0043507 | 0 | 3.59 | -31.74 | (BP) positive regulation of JUN kinase activity |
| GO:0005794 | 0 | 3.08 | -31.52 | (CC) Golgi apparatus |
| GO:0007257 | 0 | 3.08 | -31.52 | (BP) activation of JUN kinase activity |
| GO:0051098 | 0 | 3.08 | -31.52 | (BP) regulation of binding |
| GO:0030528 | 0 | 2.56 | -31.26 | (MF) transcription regulator activity |
| GO:0051090 | 0 | 2.56 | -31.26 | (BP) regulation of transcription factor activity |
| GO:0051101 | 0 | 2.56 | -31.26 | (BP) regulation of DNA binding |
| GO:0090046 | 0 | 2.56 | -31.26 | (BP) regulation of transcription regulator activity |
| GO:0004707 | 0 | 2.31 | -31.1 | (MF) MAP kinase activity |
| GO:0019207 | 0 | 2.31 | -31.1 | (MF) kinase regulator activity |
| GO:0019887 | 0 | 2.31 | -31.1 | (MF) protein kinase regulator activity |

**Table 4.9:** Most representative GO terms (according to the log-odds ratio) to classify kinase genes as neutral. %dis, %neu: Percentage of disease-associated and neutral genes annotated with a given GO term. MF: Molecular Function, BP: Biological Process, CC: Cellular Component.

| GO term | %dis | %neu | GOLOR | Description |
|---|---|---|---|---|
| GO:0048646 | 7.81 | 0 | 32.86 | (BP) anatomical structure formation involved in morphogenesis |
| GO:0007389 | 6.25 | 0 | 32.54 | (BP) pattern specification process |
| GO:0010594 | 6.25 | 0 | 32.54 | (BP) regulation of endothelial cell migration |
| GO:0050431 | 6.25 | 0 | 32.54 | (MF) transforming growth factor beta binding |
| GO:0050679 | 6.25 | 0 | 32.54 | (BP) positive regulation of epithelial cell proliferation |
| GO:0051896 | 6.25 | 0 | 32.54 | (BP) regulation of protein kinase B signaling cascade |
| GO:0051897 | 6.25 | 0 | 32.54 | (BP) positive regulation of protein kinase B signaling cascade |
| GO:0001525 | 4.69 | 0 | 32.13 | (BP) angiogenesis |
| GO:0001871 | 4.69 | 0 | 32.13 | (MF) pattern binding |
| GO:0003002 | 4.69 | 0 | 32.13 | (BP) regionalization |
| GO:0005539 | 4.69 | 0 | 32.13 | (MF) glycosaminoglycan binding |
| GO:0009952 | 4.69 | 0 | 32.13 | (BP) anterior/posterior pattern formation |
| GO:0030246 | 4.69 | 0 | 32.13 | (MF) carbohydrate binding |
| GO:0030247 | 4.69 | 0 | 32.13 | (MF) polysaccharide binding |
| GO:0032388 | 4.69 | 0 | 32.13 | (BP) positive regulation of intracellular transport |
| GO:0033158 | 4.69 | 0 | 32.13 | (BP) regulation of protein import into nucleus, translocation |
| GO:0033160 | 4.69 | 0 | 32.13 | (BP) positive regulation of protein import into nucleus, translocation |
| GO:0042306 | 4.69 | 0 | 32.13 | (BP) regulation of protein import into nucleus |
| GO:0042562 | 4.69 | 0 | 32.13 | (MF) hormone binding |
| GO:0045428 | 4.69 | 0 | 32.13 | (BP) regulation of nitric oxide biosynthetic process |
| GO:0045429 | 4.69 | 0 | 32.13 | (BP) positive regulation of nitric oxide biosynthetic process |
| GO:0048729 | 4.69 | 0 | 32.13 | (BP) tissue morphogenesis |
| GO:0051222 | 4.69 | 0 | 32.13 | (BP) positive regulation of protein transport |
| GO:0090316 | 4.69 | 0 | 32.13 | (BP) positive regulation of intracellular protein transport |

**Table 4.10:** Most representative GO terms (according to the log-odds ratio) to classify kinase genes as disease-associated. %dis, %neu: Percentage of disease-associated and neutral genes annotated with a given GO term. MF: Molecular Function, BP: Biological Process

### 4.4.6 Benchmark of the classifiers against other methods

To test the performance of our classifiers, we compared our results with those of four well-established predictors of pathogenicity: SIFT [93], SNAP [104], SNPs&GO [108] and a kinase-specific method [107]. This set of classifiers represents a wide variety of approaches and scopes: genome-wide and kinase-specific classifiers, different classification approaches (rule-based, neural networks, linear SVMs and radial basis SVMs) and a broad set of classification features.

Unfortunately, making a fair comparison of the predictive capabilities is not an easy task [91] and many technical difficulties arise. Choosing an objective testing dataset is the most difficult, especially when the datasets used in the original publications are not equivalent. Increased predictive capabilities would be expected if the testing dataset had already been presented to the classifier during the learning process. This is very likely the case for the kinase dataset, which is a strict subset of the most commonly used training dataset [136].

The results of this benchmark are shown in Table 4.11. The two genome-wide classifiers SNPs&GO and SIFT included an evaluation using the same dataset as that used to train and evaluate our own. Interestingly, when these methods were evaluated with the protein kinase dataset, performance dropped significantly compared to those reported in the original publications for a wider range of protein families. It is worth noting that this decrease in the overall performance demonstrates that the protein kinase superfamily is a challenging scenario, justifying the need for kinase-specific classifiers at the cost of scope. For a genomic-wide scenario, general classifiers such as SNAP or SNPs&GO perform better.

Our predictor generated results with the kinase dataset comparable to those obtained by the best predictor, SNPs&GO. Our method was also better than SIFT, a reference method used by many others including ours. Given that we achieved results similar to those of the best classifier is particularly interesting since some increased prediction is expected if the kinase dataset had already been presented to the classifier during the training process.

The results of our predictor are also comparable to those of the kinase-specific method proposed by Torkamani and co-workers, the only method against which a direct comparison can be made. Unfortunately, the original publication did not provide information about recall, precision and the pathogenic mutations resulting from their method. Hence, this method was only compared for accuracy and MCC. Our results are more accurate as we correctly predicted 83.29% of the cases compared to the 77% predicted by the other method. In addition, the correlation coefficient was slightly better in our case, 0.58 compared to 0.55. These results indicate that our choice of features concentrated more predictive power.

### 4.4.7 Implementation of the predictor as a web server

We implemented our pipeline to predict mutation pathogenicity in the protein kinase superfamily as a web server, KinMut, which is publicly available at `http://kinmut.bioinfo.cnio.es`. The most important features of this web server are displayed in Figure 4.23. The server displays the mutations and the SVM score for each prediction. Mutations with an SVM score greater than -0.5 are considered damaging, according to the threshold discussed above.

| Method | Scope | Accuracy (%) | Precision (%) | Recall (%) | MCC |
|---|---|---|---|---|---|
| **KinMut** | Kinase | 83.29 | 60.03 | 75.17 | 0.58 |
| SNPs&GO [108] | Kinase† | 82.28 | 62.76 | 77.45 | 0.64 |
| Torkamani [107] | Kinase | 77.00 | - | - | 0.55 |
| SIFT [93] | Kinase† | 77.63 | 37.83 | 27.88 | 0.17 |
| SNPs&GO [108] | Genome-wide | 82.00 | 83.00 | 78.00 | 0.68 |
| SNAP [104] | Genome-wide | 78.20 | 76.70 | 80.20 | - |
| SIFT [93] | Genome-wide | 68.33 | 66.11 | 56.51 | 0.35 |

**Table 4.11:** Summary of the performance of other state-of-the-art classifiers of mutations, either general or kinase-specific. Performance was measured in terms of overall accuracy, recall and the Matthews correlation coefficient. General methods with which the prediction was run with our dataset are marked with †. The remaining results for the classifiers displayed here were taken directly from their original publications

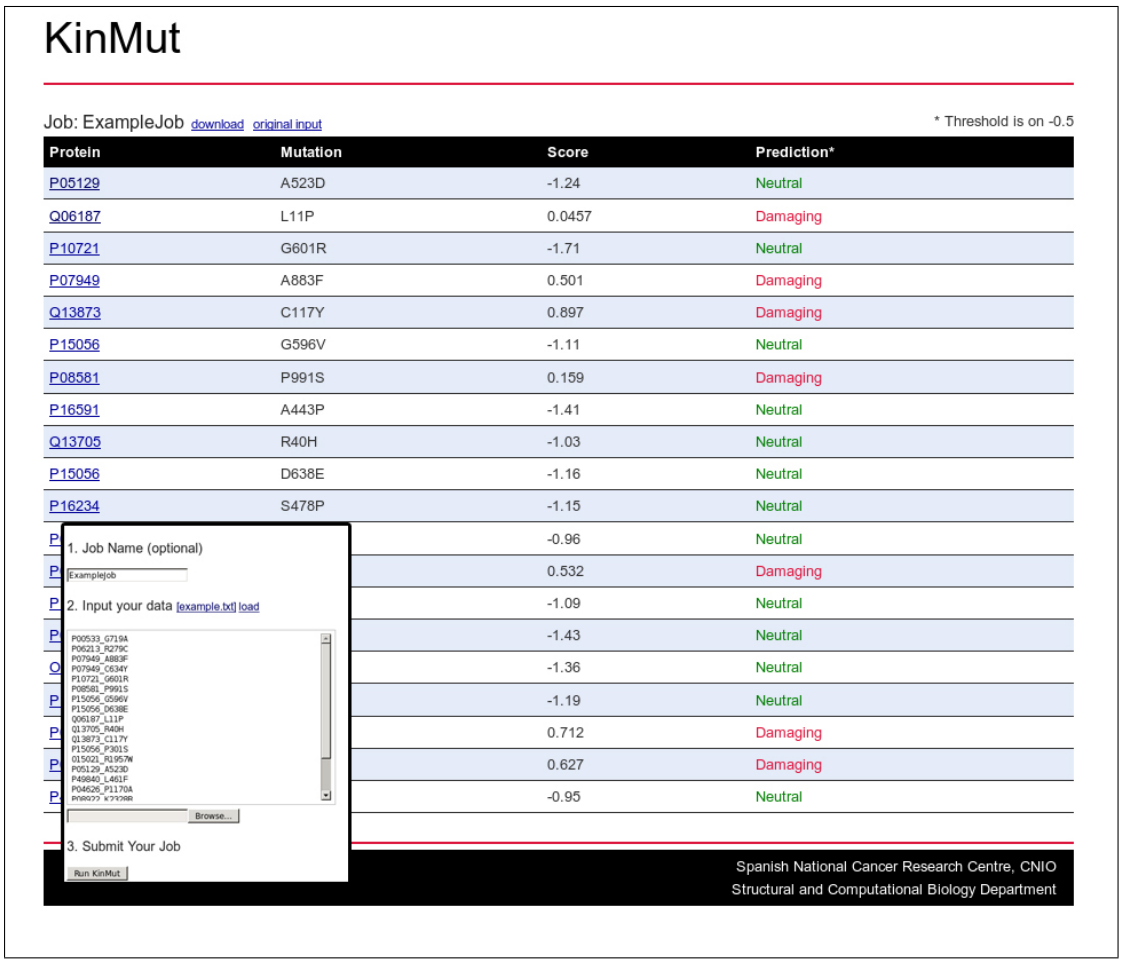



**Figure 4.23:** Schematic summary of the capabilities of KinMut.

Chapter 5

# Discussion

This thesis focuses on analyzing mutations in the human kinome, a subset of the genome chosen due to its key role in cancer and the large amount of information available on its biochemical properties. We aimed to specifically shed light on two related aspects of these mutations.

First, we wanted to characterize the mutations that disrupt the structure and function of protein kinases, focusing on the properties that contribute to differentiate pathogenic and non-pathogenic mutations. These features include the solvent accessibility, sequence and structure conservation, and the relevance of the residues for enzymatic activity or family sub-specificity.

Second, we aimed to apply the knowledge acquired to predict the consequences of mutations for human disease, and cancer in particular, using the properties associated with the mutations and the corresponding phenotypes as the input.

We approached these scientific questions from the perspective of Computational Biology. This discipline provides us with the methodology to identify mutations and their properties, link them to annotations from heterogeneous information sources, classify the mutations based on the features selected and according to the information available, and use this classification to predict the effects of newly discovered mutations.

## 5.1 3Dsim: Structural implications of mutations

One of the initial steps taken in this thesis was to implement the 3DSim method [143] to interactively map single amino acid polymorphisms onto protein structures. In addition, 3DSim summarizes several sources of information into a single repository and it displays annotations about the mutations, such as wild-type and mutant sequences, predicted structural implications of the mutations, and their characterization as neutral or pathogenic according to these predictions. To enhance the potential to use this information, the system also provides links to the original repositories: UniProt [45], Gene3D [120], SAAPdb [74], CATH [121] and Modbase [144] among many others.

Successful attempts to map and display mutations, along with their annotations onto a protein structure, had been completed previously (see Table 2 in Uzun *et al.* (2007) [167] for a detailed review). However, the method discussed here presents several unique and interesting features that are not available in earlier implementations.

First, the similar treatment of SNPs, and the less common and more harmful pathogenic deviations (PDs), allows the users to inspect and compare both kinds of mutations through the same interface, including explanatory annotations where available.

Second, the localization of mutations within the CATH hierarchy allows users to query and explore the distribution of the mutations at different levels of the structural classification.

Third, it is possible to transfer mutations and annotations from protein sequences to related three-dimensional structures, even when no structures are available. In such cases, we use CATH [121] and Gene3D [120] to assign a representative structure that can be used as a proxy for the real structure. The possibility of using an alternative structure as a representative of the mutated sequence has been successfully employed by other methods, such as those of LS-SNP [102], stSNP [167] and Modbase [144], although with slight differences in terms of implementation. The main difference is that the earlier methods used a structure calculated with MODELLER [156], whereas in our implementation the representative structure corresponds to one of the manually curated domain representatives in CATH that is chosen by a sequence-similarity search (see the Methods section for a detailed description of our algorithm).

Finally, the availability of data via web services and databases enables users to include this information efficiently into their own analyses. These facilities allow the independent integration of our data into any other pipeline or workflow.

## 5.2 Automatic extraction of human kinase mutations from the literature

To test the performance of our methodology, we compared the mutations extracted by our pipeline to those from the most frequently used annotation databases and a compilation of data from genotyping studies. Through this comparison, we concluded that our pipeline increased the total number of database records by extracting mutation mentions from the literature. Moreover, automated mutation extraction systems were demonstrated to be a valuable resource to assist manual curation by providing direct pointers to sentences of mutation evidence that can be rapidly examined by experts.

One such example is the Y845F mutation in the epidermal growth factor receptor (EGFR). Y845F is one of the 32 mutations in this protein that was extracted from the literature and that was not present in the databases queried. We recovered a number of sentences mentioning this mutation: '*Furthermore, transient expression of a Y845F variant EGFR in murine fibroblasts resulted in an ablation of EGF-induced DNA synthesis to non-stimulated levels.*' (PubMed:10075741); '*Stably transfected B82L cells with a point mutation of the EGFR at Tyr-845 (B82L-Y845F) exhibited only basal Ras activity following exposure to* $Zn^{2+}$' (PubMed:11983694); and '*In contrast, LPA-elicited DNA synthesis and migration were augmented in cells expressing EGFR, EGFR(K721A), or EGFR(Y845F), but not EGFR(Y5F), although the PDGF responses were indistinguishable*' (PubMed:15364923). The information retrieved is sufficient to suggest to the experts that Tyr-845 might be involved in DNA synthesis triggered by EGF binding to its receptor. In this case, it is also important to provide the expert with additional information regarding this residue, for example, that it actually maps to position 869 in the protein sequence due to a signal peptide that was not considered by the authors of the publication.

In the Introduction to this thesis, we presented other methods for the automatic extraction of mutations from the literature. The added value of our system in comparison to those methods is the inclusion of a multi-layer pipeline to corroborate the presence of the extracted mutations in the protein sequences. This validation pipeline not only includes a basic comparison of the mutations to the reference sequence in UniProt but also, more intricate strategies such as a sliding window algorithm that searches for a pattern of mutations along the sequences instead of at exact positions, and the consideration of alternative starting points due to signal peptides or methionine cleavage (this pipeline is described in detail in the Results section).

We retrieved approximately one third of the mutations present in the most commonly used databases and genotyping studies. It is most likely that the remaining mutations were not obtained as there is insufficient evidence in the text, probably due to:

1. *Missing accessibility to the original reference, full-text articles or additional materials.* Little can be done from our side with respect to this issue. This is especially evident in older publications that are not necessarily digitalized, although there are currently efforts to digitalize the literature that only exists as print copies. Nevertheless, this issue will continue to be unsolved unless the publishing houses make their archives openly accessible and, from a technical point of view, the formats and repositories comply with the needs of the automatic extraction pipelines.

2. *General methodological limitations of mutation mention extraction*, such as those associated with MutationFinder [90] in our case. This is especially relevant in terms of the number of mutations retrieved (recall) and the reduction in the number of incorrect mutation mentions (false positives). We expect that better mutation extraction methods will be available in the near future.

3. *Limitations in protein normalization and mutation to sequence annotation.* We plan to improve our extraction pipeline in several ways. Thus, future implementations of our system will very likely include mutation-protein proximity analysis, examination of species and organism source ambiguity, and analysis of the probability of finding a given mutation within a target sequence by considering the observed residue composition of the proteins, and particularly kinases.

4. *Improvement of the sequence validation pipeline.* Our system currently includes a multi-step pipeline to validate the mutation mentions extracted and to reduce the number of false positives. Basically, we corroborate the existence of the residues at a given position of the reference sequence in UniProt [45] and correct for alternative starting points, i.e., due to the presence of the initial methionine or signal peptides. However, multiple improvements to this pipeline can be made, the most straightforward of which is the consideration of alternative isoforms. As different splicing isoforms skip or include exons, the numbering of the positions change downstream of the splicing event. These alternative counting possibilities need to be included in further developments.

The scope of this work is not only limited the identification of mutations for which evidence in the literature exists but there are no records in the databases. Even more interestingly, the work presented here proves that the extraction of mutations from full-text articles is feasible and that it is a powerful resource to characterize and annotate mutations. Indeed, one of the strengths of the system is the possibility of linking mutations with supporting experimental evidence from the literature. This additional information usually includes crucial data such as experimental conditions, organisms, evidence of association with disease or phenotype and,

in the best case scenario, a description of the underlying biochemical mechanisms affected. This background information is very valuable and consequently, we think that systems such as personalized medicine platforms or pathogenicity prediction pipelines may benefit from the information contained in our system.

Although we focused on the extraction of mutations within the protein kinase superfamily, nothing impedes the application of the described methodology to the whole set of full-text articles in PubMed. Despite representing a computationally expensive task, the huge amount of information that would be gathered greatly justifies this effort. Nevertheless, this approach is definitely not intended to substitute human curation and expert validation but rather, it is a complement and guide, helping prioritize the efforts of manual curators. See Leitner *et al.*, 2010 and Krallinger *et al.*, 2011 for a detailed description of the efforts to integrate text mining in database pipelines and manual curation [81, 168].

## 5.3 Preferential distribution of disease-associated kinase mutations

We compared here the distribution of point mutations in protein kinases in order to characterize the structural and functional consequences of the different types of mutations. In particular, we focused on the protein kinase domain. The human kinome is particularly amenable for this type of analysis because it represents one of the best characterized protein superfamilies [40, 43, 46, 55, 169, 170], and because a growing number of drugs already target members of the protein kinase family [36, 37]. Therefore, many efforts have been made to crystallize kinase protein structures and a large number of mutations are currently available as discussed above [14, 27, 71, 74, 75, 171]. Furthermore, the underlying biochemical mechanisms have also been characterized for a growing subset of these mutations.

To cover all aspects of kinase mutations and their relationship to disease, we analyzed somatic and germline mutations [154, 155]. Germline mutations are inherited from parents and transmitted to offspring, while somatic mutations are acquired variations that differ from the progenitors' fertilized eggs [9]. It is important to understand this difference and the implication it has in understanding the link between kinase mutations and disease. Hence, while germline mutations cause inherited diseases [70] and kinasopathies [171], cancers typically arise due to somatically acquired mutations [9, 14, 16].

We analyzed a number of features relevant to protein kinases, including: (a) sequence conservation as an indicator of evolutionary important regions; (b) localization in the buried core of the protein; (c) organization of mutations with respect to regions relevant for enzymatic activity; and (d) specificity-determining positions.

We first analyzed the distribution of somatic mutations with respect to these features [154] and classified the mutations into two different types: drivers and passengers. Driver mutations correspond to those causally implicated in oncogenesis as they confer growth advantage on the cancer cell. These mutations are positively selected in the microenvironment of the tissue in which the cancer arises. By contrast, passengers do not confer any clonal growth advantage or contribute to cancer development.

Driver and passenger somatic mutations display clear differences in their distributions in the protein kinases [154]. Driver mutations tend to occur close to regions important for protein function and structure, such as the ATP- and substrate-binding sites, residues that

confer subfamily specificity, conserved residues and the hydrophobic core of the protein. This distribution is consistent with their association with disease and their putative role in the development of tumors. By contrast, passenger mutations do not occur in buried or conserved residues and they are not usually clustered around functionally relevant sites. The neutral role of these mutations is attributable to their localization in these regions with mild effects.

In a related experiment, we characterized the preferential distribution of germline mutations with respect to functional and structural features, and according to their propensity to be involved in disease. Thus, we classified the mutations as pathogenic deviations or neutral polymorphisms. There were significant differences between the two types of germline mutations in terms of conservation, accessibility and distance to the residues forming the ATP-binding site or the regions that provide family sub-specificity. In fact, the behavior of these mutations mimicked that previously discussed for driver and passenger mutations (Table 5.1).

| **Feature** | **Somatic** | **Germline** |
|---|---|---|
| Sequence conservation, Shannon's entropy [125] | Drivers | PDs (Slightly) |
| Sequence conservation, AL2CO [126] | Drivers | PDs |
| Accessibility | No difference | PDs |
| Catalytic site, FireDB [127] | Drivers | PDs |
| Catalytic site, Knight [46] | Drivers | PDs |
| Tree determinants | Drivers | PDs |

**Table 5.1:** Summary of the preferential distribution of kinase mutations with respect to the features analyzed. Disease-associated mutations, drivers and pathogenic deviations (PDs) tend to occur closer to the important regions of the protein than their neutral counterparts.

Interestingly, despite the similar location of the mutations, both datasets are non-redundant and indeed, only four mutations are common to the driver and pathogenic datasets, all of which are found in the B-raf proto-oncogene. This is consistent with the very different nature (germline and somatic) of the mutations in each of the disease-prone datasets.

The results presented here confirm the functional and structural relevance of the disease-prone sets of mutations: the germline pathogenic deviations and somatic driver mutations. The fact that driver mutations behave similarly to other pathogenic mutations could indirectly support the categorization of mutations into drivers and passengers, and enable them to be used as a proxy to study the involvement of somatic mutations in cancer biology.

Moreover, these results constitute a necessary step towards developing a predictor that uses machine learning techniques. Some of the features described here would be able to predict the pathogenicity of any novel kinase mutation. This predictor has been introduced in the Results section and its implications will be addressed below.

The trends observed in our experiments for both germline and somatic mutations are consistent with those defined previously [172]. In those studies, the authors classified inherited germline mutations with a customized machine learning approach and ranked the features that contributed most to the classification. In addition, the mutations were mapped onto structural elements of the protein kinase domain. The conclusion of these studies was that mutations

predicted to be disease-associated tended to map to regulatory and substrate-binding regions. The results we obtained suggest a similar tendency. Thus, both sets of results demonstrate the relationship between disease-associated amino acid changes and key regions of protein structure and function, although from slightly different perspectives.

The analyses performed here have some limitations. The experimental information available is currently very limited and the majority is derived from high-throughput genetic studies that lack further validation. The biochemical characterization of mutations cannot keep up with the pace of the techniques to discover mutations, since the study of a single mutation out of the millions identified requires an enormous amount of effort, time and resources. This is especially difficult for protein kinases, which are regulated by tightly controlled cascades where changes in one protein can elicit phenotypes that are often identified as disease or altered states associated with downstream or negative signaling components. This is the case, for example, of mutations in the protein tyrosine phosphatase that activates the kinase signaling pathway in Noonan syndrome [173].

Moreover, at the functional level there is an important difference between *gain-of-function* and *loss-of-function* mutations. Changes in function occur at various levels, from the protein level where the changes in a residue directly perturb protein function, to the system level where the mutations can affect the whole cell and/or organism. For example, a mutation that perturbs the function of a protein kinase that is part of a signaling cascade might disrupt the activity of the proteins downstream in the same cascade, and in turn, the cellular functions dependent on that signaling event. In this case, the predicting the pathogenicity of a mutation based on the mechanistic understanding of the function of the signaling cascade requires a detailed knowledge of the full system, and the application of an adequate system simulation strategy. Even if progress in this area of Systems Biology is made, reliable methods connecting mutations and system response are still not available and they are beyond the scope of this thesis.

A gain-of-function mutation constitutively increases kinase activity, sometimes leading to unrestrained signaling, and it may trigger oncogenesis or cause rare inherited dominant phenotypes. This is the case of the somatic mutation V600E in B-raf that is associated with non-Hodgkin lymphoma, colorectal cancer, malignant melanoma, thyroid carcinoma and lung carcinoma [148, 174, 175]. By contrast, loss-of-function mutations may lead to a decline in cell signaling, which can in turn affect cell growth and normal tissue development. This is exemplified by mutations in the proto-oncogene RET, which are associated with Hirschsprung disease, a type of congenital intestinal aganglionosis affecting the enteric ganglia in the intestinal tract [176, 177]. Another interesting example of a loss-of-function mutation is the B-raf kinase, where there is a link between the deactivating mutations E501G and G596V, and CFC syndrome [175]. Please refer to Lahiry *et al.* (2010) [171] for an interesting review on this subject. In addition to their important relationship with disease, the over-activation or over-repression of kinase activity could provide insights that might help characterize the biochemical mechanisms governing kinase activity.

Although the results presented here focus on the mutations in the protein kinase superfamily, the methodology described here to analyze the distribution of mutations according to a fixed set of features can be applied to any other protein family of interest. Indeed, it would be an interesting forthcoming project to analyze other protein families at a genome-wide scale to corroborate the extent to which the trends observed for the protein kinase subfamily can be extrapolated to the whole human genome.

## 5.4 Features that determine the pathogenicity of kinase mutations

The majority of protein kinase mutations are tolerated without producing significant effects, while only a small subset is causally associated with disease. In cases where the mechanistic interpretation of the consequences of the mutations on phenotypes is not feasible, current research aims to identify correlations between mutations and human diseases, particularly cancer. Current advances in automatic machine learning enable the rules to be generalized based on prior observations, which can then be used to assess the probability of a mutation being harmful. It is in this context that we propose KinMut, a computational method that predicts the impact of mutations on protein kinase function from basic sequence information.

Based on the knowledge acquired through our experiments to study the preferential distribution of neutral and pathogenic kinase mutations with respect to certain types of residues, we linked a number of sequence-derived features to disease-associated kinase mutations, including: (a) membership of a KinBase group and Gene Ontology terms at the gene level; (b) the occurrence of the mutation inside a PFAM domain at the domain level; and (c) several other properties including amino acid type, functional annotations from SwissProt and FireDB, and specificity-determining positions at the residue level. We examined the independent significance of these properties and their combination using a support vector machine (SVM).

Good features are those that can discriminate between pathogenic and neutral mutations. To decide which of these are the most adequate, smart feature selection must have a wide collection of mutations from both categories (pathogenic and neutral) and a representative variety of possible values within both categories. The underlying assumption is that the collection is large enough and includes a sufficient representation of the variation in each of the classes. If the collection is too small, not representative of the pathogenic or neutral classes, or too contaminated with mislabeled mutations (truly pathogenic mutations labeled as neutral or *vice versa*), we will not be able to classify the mutations properly.

Our classifier relies solely on sequence-based features. We are aware that a myriad of available protein features exists and that it is unfeasible to integrate all of them and all their combinations into a classifier. For instance, other methods include structure-based features in their calculations [99, 102, 103, 141, 178, 179]. It is evident that protein structures retains a huge amount of relevant information. Concomitantly, the protein space of the prediction is limited to only those proteins for which a structure has been properly determined, or at least for which a reliable enough model can be obtained. For a broader coverage, we designed our classifier around sequence-based features and demonstrated its ability to achieve a reasonable level of accuracy. It is worth noting that, in the future, when the gaps in the structure-space are complete, the detection of pathogenic mutations will very likely benefit from the inclusion of structure-based features without risking the scope of the prediction.

The selection of the training dataset is crucial. Previous methods have used a variety of training sets to construct their pathogenic and neutral catalogues. Disease-associated mutations are usually collected from databases such as the SwissProt variant pages [71, 99, 105, 108, 109, 178], OMIM [107, 180], COSMIC [27, 105], KinMutBase [75, 107], HGMD [95, 103, 107, 164, 181], or from clinical or functional studies. Generally, these mutations refer to Mendelian or monogenic diseases [91]. The neutral dataset is usually gathered from the SwissProt variant pages as well, but also from dbSNP [67, 95, 105, 107, 164], HapMap [69, 164], HGVbase [98, 182], or generated artificially from the equivalent amino acids in closely related species [164, 178, 183]. Many

early predictors undertook a different approach and used the results of saturation mutagenesis experiments in bacterial and viral proteins [93, 184]. Another approach is to use all the available information about possible deleterious/neutral amino acid substitutions (human disease databases and systematic functional assays among others) in order to construct large benchmark sets, such as differences between two closely-related functional enzymes that can be considered neutral [104]. In a recent benchmarking of several training datasets [136], it was concluded that the SwissProt variant pages provided the most accurate results to predict the pathogenic effect of human non-synonymous mutations. Consequently, we derived our datasets from this repository.

Nevertheless, several issues have to be considered in data analysis. Common to any of these benchmark sets is the underlying assumption that a cause-and-effect relationship exists between the mutation and the elicited phenotype, which may be debatable. In addition, the pathogenic dataset assumes a simple one-to-one relationship between mutation and disease. In complex diseases such as cancer, the reality is that the mutations act as a contributing rather than a causal factor.

## 5.5 Performance of the classifier and the benefits of family-specific prediction

Our KinMut predictor achieves a level of performance similar to that obtained by the best predictors, SNPs&GO [108] and Torkamani's kinase-specific [107] method, and it outperforms other reference methods such as SIFT [93] when evaluated within the framework of the kinase dataset.

Interestingly, we achieved results comparable to the best classifier, SNPs&GO, whose capability we probably overestimated due to a technical artifact. SNPs&GO has been trained with all the mutations in UniProt, which are very likely to be included in our subset of kinase mutations. When this set was presented to SNPs&GO for classification in order to compare its performance with that of KinMut, SNPs&GO was given some advantage since the classifier had already been presented with the mutations. Thus, it is not possible to compare the classifiers without training the predictor from scratch without the kinase mutations, which would nevertheless yield an unfair comparison as well. Therefore, we assume this artifact to be acceptable for our qualitative analysis.

Probably, the similarity in the performance is given by the use of GO terms since both methods - even though they differ in their scope and implementation - benefit from functional information encoded as GO terms at the gene level, which is the most discriminative feature for our classifier.

Our predictor performs beyond the capabilities of the only method against which an utterly fair comparison can be conducted, Torkamani's kinase-specific predictor [172], at least in terms of accuracy and correlation. Unfortunately, the authors of this method did not provide information about its recall, precision and output, to enable a better comparison to have been made. Interestingly, Torkamani's and our classifiers share several properties: amino acid types; kinase group membership, which the authors state to be critical for classification; biochemical properties such as the Kyte-Doolittle hydrophobicity index; and evolutionary conservation. In spite of these similarities, Torkamani's method does not benefit from intra-family specificity positions or from GO annotations, which we have shown to be crucial for prediction (see above). This might have caused the differences in performance observed.

During the writing of this thesis, another method was published that takes into account specificity-determining positions [109]. Hence, we are looking forward to evaluating the performance of this classifier with our dataset in order to compare different implementations with a very similar concept.

Current genome-wide predictors of mutation pathogenicity perform well on average, probably because they can use the huge amount of mutation data available. However, most of these predictors only exploit a small proportion of the possible characteristics for the sake of coverage, the subset of features that could be generalized to the entire range of protein families in the human proteome, which constitutes an intrinsic limitation. By contrast, family-specific predictors, such as the method presented here for the protein kinase superfamily, can overcome this limitation and benefit from features that apply only to the protein family of interest. These family-specific features might capture aspects of pathogenicity that are unique to that given protein family. We explored the basis of using kinase-specific features, such as kinase group membership, annotation with certain GO terms and the presence of determined Pfam domains, which are relevant for predicting pathogenicity in the protein kinase superfamily. Accordingly, the performance of genome-wide methods decreases when they are confronted with the set of kinase mutations.

Indeed, the family-specific nature of our method allowed us to explore features that are unique to the protein kinase superfamily, retaining valuable information on mutation pathogenicity. In our case, membership to a particular kinase group and the occurrence of mutations in the catalytic protein kinase domain were important features that are unique to the protein kinase superfamily. This is in full agreement with previous observations that reached similar conclusions [107, 172].

The results provided here reinforce the idea that for well-studied families like the kinase superfamily, family-specific classifiers can use unique features that are only valid in the context of this specific superfamily, thereby improving performance over general purpose methods. Regarding the dissimilarities between the different branches of the kinase phylogenetic tree, we demonstrated that more accurate results were obtained for groups with sufficient data that allowed the classifier to learn to weight the importance of the individual contribution of the features precisely. Moreover, this group membership was one of the most triggering features of our classification.

There are groups in which very few (or even no) pathogenic mutations have been described and as such, in these cases group membership is a powerful means to predict neutral mutations. However, since the mutational landscape is far from complete, we cannot discern whether this is a reliable scenario (where these kinase groups do not elicit pathogenicity) or rather an artifact due to a gap in our current knowledge that will be filled when new mutations are discovered.

Indeed, the uneven, heterogenic, distribution of experimental evidence regarding the different kinase groups does not only affect the number of mutations discovered but also, the quality and thoroughness of features such as GO or UniProt annotations, which is very likely to influence the predictive capacity of our system.

In the near future, ongoing genomic projects will help us understand the links between mutations in all the kinase groups and disease, thereby boosting the capability of kinase-

specific prediction methods beyond the limits of current highly populated groups.

## 5.6 Prioritization of pathogenic mutations for experimental characterization

Our method was not conceived for direct use in a clinical situation. Indeed, currently no *in silico* method can achieve the required accuracy to be a substitute for laboratory experiments.

Even if the method we have developed performs at a level similar to that of the best methods available, the results are still of limited direct use. Indeed, the most direct practical use in our case is as a component of a system that ranks mutations generated in cancer genome studies for further experimental validation.

Additionally, the development of the system opens new possibilities to study the relative importance of the features used by the classifier, information that could help to investigate the underlying molecular events linking mutations, biochemical/cellular events and pathogenesis.

A limitation common to all current predictors of pathogenicity, including ours, is that they only provide results for missense variations. New methods are necessary to evaluate the functional impact of synonymous mutations and small insertions/deletions (commonly referred to as indels). Methods that determine the impact of mutations occurring outside of the coding fraction of the genome are also needed. Although in these cases no aberrant protein product is expressed, promoter regions, transcription factors and splicing sites may still be affected, along with other important non-coding regions.

## 5.7 Pathogenicity prediction and personalized medicine

One of the most interesting applications of the methodology presented here is its use in prioritizing mutations for personalized medicine platforms.

The purpose of personalized medicine is to provide each patient customized treatment that takes into account the predicted response to drug therapy, instead of administering a universal treatment based on trends observed in the whole population. Personalized prediction of drug response is founded on the sequencing of the patient's genome. Our group recently developed a prototype to help clinicians choose personalized cancer treatment according to the patient's genomic profile. The objective of this platform is two-fold: prioritize somatic mutations in the genome of the patient according to their role in tumor onset and progression; and suggest candidate drugs for personalized treatment. Our predictor of pathogenicity might represent a valuable tool for prioritizing somatic mutations.

The pipeline of the initial personalized medicine prototype can be described as follows. For each somatic mutation obtained in a full-exon sequencing of the patient's genome, the system automatically collects information about mutated genes, the corresponding proteins and the pathways involved. The application also predicts the potential functional effect of the mutations using several state-of-the-art predictors of the mutation pathogenicity such as SIFT [93], Polyphen II [99] and SNPs&GO [108].

It is evident that other information sources will be added in the future to further de-

velop this analysis platform. In particular, the development of the KinMut predictor described here fits perfectly with the need to predict the consequences of mutations in protein kinases, as they are frequently mutated in cancer [14, 27, 71, 74, 75, 171] and they are common clinical targets of a number of cancer drugs [35, 36].

## 5.8 Future developments and perspectives

In addition to its incorporation into a personalized medicine platform, a number of other extensions and improvements of the predictor presented here are envisaged.

From a technical point of view, the most straightforward future task regards the maintenance of the server. The performance of the classifier correlates with the amount of input information and the number of experimentally characterized mutations. Therefore, the classifier would need to be re-trained and re-evaluated with each new release of the mutation databases. Thus, an obvious future objective is to design a protocol for the automatic maintenance of the predictor whenever mutation databases are updated.

From a conceptual perspective, future plans include the introduction of structure-based features to complement the sequence-based features used in our prediction pipeline, such as protein flexibility [104, 107], structural conservation [185] and allosteric regulation [50, 51, 55, 164]. The 3DSim method [143] introduced in this thesis will be an important part of this development, since it will provide the means to map the position of the mutations onto corresponding protein structures. Hopefully in the near future, more structures, including reliable models, will be available. With the completion of the structure-space, we will be able to include new features that can be scaled up to a large number of structures, and study more intricate processes, such as the implication of mutations in protein-protein interaction or drug binding.

We are also interested in extending the work started here with the protein kinase superfamily towards other protein families of interest in cancer biology, such as phosphatases, caspases and small GTPases.

With the growing number of families to be analyzed, this project would require the development of automatic pipelines to fulfill the tasks that were completed manually here. Thus, different kinds of pipelines would be required.

First, we would need to retrieve the mutations and classify them as pathogenic or neutral. The text mining system for the extraction of kinase mutations that we discussed previously can be adapted to retrieve the necessary information. Detecting mutation mentions and mapping them onto proteins outside the protein kinase superfamily would not be a problem for the system as it is. However, the detection of the classification of the proteins would require a development similar to that implemented in our previous work, where a machine learning approach was used to detect the natural or mutagenic character of the kinase mutations (data not shown, [150]).

Second, we need to calculate the support vectors that describe each of the mutations in the terms of the selected features. These include pipelines for aligning sequences, calculating conservation, extracting catalytic sites from FireDB, gathering annotations from UniProt and calculating specificity-determining positions, among the general features. This can be achieved with the current development of the system and no additional issues would be expected. However, one of the main difficulties we will face is the systematic selection of family-specific features for mutation classification. As we discussed previously, one of the strongest aspects of

our kinase-specific predictor is the presence of customized features that are only valid in the context of the protein kinase superfamily. Thus, the automatic selection of specific features still needs to be solved.

Finally, to evaluate which features would be the most informative, we would need an automatic self-assessing system, and since the number of features is relatively small, an exhaustive search will be sufficient. If the number of possibilities grows beyond reasonable limits, other implementations can be tested, such as the Monte Carlo approach where the individual contribution of inserting a given feature is evaluated.

In the future, we plan to cover the entire protein-space and train classifiers for the remaining protein families, integrating and validating these information sources in the automatic machine learning framework initially developed for the protein kinase superfamily. With the completion of the family-space, we will address whether the trend of disease-associated mutations to localize near important residues is general or a particularity of the protein kinase superfamily, and whether the definition of driver somatic mutations is valid for other protein families.

Chapter 6

# Conclusions

1. We have developed a system, 3Dsim, to *transfer mutations and annotations from sequence coordinates to the corresponding three-dimensional protein structure*. The system exploits the capabilities of CATH and Gene3D to assign a representative structure to the query sequences.

2. We have implemented an automatic pipeline that can *extract mentions of mutations in protein kinases from abstracts and full-text articles and map them onto corresponding protein sequences*. Almost one half of these mutations were not previously recorded in public databases. The extracted mutations are a useful information repository.

3. We have demonstrated that automatic extraction pipelines are valuable for *linking protein kinase mutations and their mentions in the literature*. These pipelines are useful in assisting manual curation of mutations, e.g., by providing experimental conditions, functional evidence and the associated phenotypes.

4. Differences exist in the distribution of *somatic mutations with respect to important regions in the structure of human protein kinases*. Potential cancer-associated mutations (drivers) tend to occur near conserved and functionally relevant regions, such as the ATP-binding pocket and subfamily specificity positions relevant for effector binding.

5. *Germline mutations associated with disease* (pathogenic deviations) tend to be located close to conserved solvent-inaccessible regions, the ATP-binding pocket and family specificity-determining positions.

6. Somatic driver and germline pathogenic mutations have a similar distribution in important regions of protein kinase structures. This might be interpreted as supporting the classification of *pathogenic somatic mutations as cancer drivers*, since their distribution is similar to mutations known to be able to elicit disease.

7. We have implemented an automatic system to *predict the pathogenicity of mutations in the protein kinase superfamily*. This system integrates general and family-specific information, including sequence-based features and experimentally-derived functional annotations. The results support the use of family-specific predictors and justifies its extension to other protein families.

# APPENDICES

# Resumen y Conclusiones

## Resumen

Las superfamilia de las proteína quinasas constituye una prometedora diana farmacológica debido a que estas proteínas están involucradas en gran número de procesos tumorigénicos tales como evasión del sistema inmune, proliferación, anti-apoptosis, metástasis y angiogénesis. La mayoría de las mutaciones descritas en proteína quinasas son perfectamente toleradas y no se conoce que sean causa de enfermedad o fenotipo adverso alguno. Consecuentemente, podemos suponer que son neutras para la función y estructura de las proteínas. Por el contrario, para un reducido número de mutaciones es patente su implicación directa en alguna enfermedad.

Los mecanismos mediante los cuales las mutaciones dan lugar a fenotipos adversos han sido estudiados y caracterizados bioquímicamente en muchos de estos casos y algunas de las relaciones causa-efecto están ampliamente estudiadas en la actualidad. Sin embargo, la caracterización en términos bioquímicos de las mutaciones requiere semejante cantidad de recursos que no es factible en términos prácticos mantener el ritmo impuesto por las tecnologías de alto rendimiento que existen en la actualidad para la identificación de nuevas mutaciones.

Es necesario, por tanto, ampliar nuestro conocimiento sobre los mecanismos mediante los cuales las mutaciones alteran la función de las proteínas y dan lugar a enfermedad. Esto nos permitirá desarrollar protocolos mas eficientes para la caracterización y priorización de mutaciones, de tal forma que los esfuerzos de la comunidad científica se orienten hacia aquellas mutaciones que más probablemente jueguen un papel causal en el desarrollo de enfermedades.

El objetivo de esta Tesis Doctoral es ampliar nuestro conocimiento acerca de los mecanismos mediante los cuales las mutaciones patogénicas interfieren con el normal funcionamiento de las proteína quinasas, dando lugar a enfermedad así como desarrollar una plataforma para identificar aquellas mutaciones en proteína quinasas que presenten una alta probabilidad de estar involucradas en enfermedad. Al centrar nuestros esfuerzos en la superfamilia de las proteína quinasas podemos utilizar métodos e ideas especícos para su organizacion y su evolución en familias de proteínas.

Estas interesantes cuestiones científicas se tratan en esta Tesis Doctoral desde la perspectiva de la Biología Computacional. Dicha disciplina proporciona el entorno necesario para el análisis integrado de grandes cantidades de información provenientes de diferentes fuentes. Además, los avances en bioestadística y en aprendizaje automático posibilitan la generalización de reglas basadas en observación del comportamiento de las mutaciones previamente caracterizas. Dichas reglas permiten la estimación teórica de la probabilidad de que una mutación dada altere la función de la proteína y resulte desencadentante de alguna tipo de enfermedad.

## Conclusiones

1. Hemos desarrollado un sistema, 3DSim, que permite la *transferencia de mutaciones y sus anotaciones desde las secuencias a las estructuras tridimensionales de las proteínas.* El sistema hace uso de CATH y Gene3D para asignar estructuras representativas a las secuencias si es necesario.

2. Hemos implementado un protocolo automático para la *extracción de mutaciones en proteína quinasas mencionadas en artículos científicos y para mapear dichas menciones a sus secuencias proteicas correspondientes.* Casi la mitad de las mutaciones recuperadas de esta forma no estaban presentes en las bases de datos publicas y constituyen, por tanto, un nuevo repositorio de utilidad reseable.

3. Hemos demostrado que la extracción automática de mutaciones en el ámbito de las proteína quinasas constituye un recurso válido para la *asociación de mutaciones con sus menciones en la literatura.* Este tipo de herramientas facilitan la curación manual de las mutaciones y proporcionan información adicional como, por ejemplo, las condiciones experimentales, las funciones de las proteínas o los fenotipos asociados.

4. Existen diferencias en la distribución de las *mutaciones somáticas con respecto a regiones importantes en la estructura de las proteína quinasas humanas.* Las mutaciones potencialmente asociadas con cáncer (drivers) tienden a localizarse cerca de posiciones conservadas y regiones relevantes para la función como el sitio de unión de ATP, las regiones de unión a efectores y ligandos o los residuos que confieren especificidad dentro de las subfamilias de quinasas. Esta asociación es coherente con el papel potencial de estas mutaciones en los primeros estadios del desarrollo de tumores.

5. Estas diferencias también existen en la distribución de las *mutaciones germinales con respecto a las regiones importantes en la estructura de las quinasas humanas.* Las mutaciones patogénicas, aquellas para las que existe una validación experimental de su asociación a enfermedad, tienden a encontrarse cerca de regiones conservadas, en regiones inaccesibles para el solvente, cerca del sitio de unión de ATP o en posiciones importantes para la especificidad de subfamilias o la unión de efectores.

6. Tanto *mutaciones somáticas como germinales presentan una distribución parecida con respecto a las posiciones importantes para la estructura del dominio quinasa.* Esta observación se puede interpretar como un argumento a favor de la controvertida categorización de las mutaciones somáticas en drivers y passengers y su uso como un modelo para el estudio del papel de las mutaciones somáticas en la biología del cáncer.

7. Hemos implementado un sistema automático para la predicción de la *patogenicidad de las mutaciones en la superfamilia de las proteína quinasas.* Este sistema integra tanto información general como específica de familia. Esta información incluye propiedades basadas en secuencia y anotaciones funcionales obtenidas experimentalmente. *Los resultados presentados avalan el desarrollo de predictores específicos de familia y la extensión de este trabajo a otras familias mas allá de las proteínas quinasas.*

# References


[1] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–51.

[2] Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27: 234–6.

[3] Collins FS, Brooks LD, Chakravarti A (1998) A dna polymorphism discovery resource for research on human genetic variation. Genome Res 8: 1229–31.

[4] Consortium GP, Durbin RM, Abecasis GR, Altshuler DL, Auton A, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–73.

[5] Cooper DN, Smith BA, Cooke HJ, Niemann S, Schmidtke J (1985) An estimate of unique dna sequence heterozygosity in the human genome. Hum Genet 69: 201–5.

[6] Kwok PY, Deng Q, Zakeri H, Taylor SL, Nickerson DA (1996) Increasing the information content of sts-based genome maps: identifying polymorphisms in mapped stss. Genomics 31: 123–6.

[7] Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY (1998) Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. Genome Res 8: 748–54.

[8] Gibson G, Muse SV (2001) A primer of genomic science. Sinauer Associates Inc : 241–98.

[9] Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458: 719–24.

[10] Bardelli A, Parsons DW, Silliman N, Ptak J, Szabo S, et al. (2003) Mutational analysis of the tyrosine kinome in colorectal cancers. Science 300: 949.

[11] Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133–41.

[12] Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, et al. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40: 722–9.

[13] Baudot A, Real F, Izarzugaza J, Valencia A (2009) From cancer genomes to cancer models: bridging the gaps. EMBO Rep .

[14] Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, et al. (2007) Patterns of somatic mutation in human cancer genomes. Nature 446: 153–8.

[15] Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, et al. (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314: 268–74.

[16] Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, et al. (2007) The genomic landscapes of human breast and colorectal cancers. Science 318: 1108–13.

[17] Jones S, Zhang X, Parsons DW, Lin JCH, Leary RJ, et al. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321: 1801–6.

[18] Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, et al. (2008) An integrated genomic analysis of human glioblastoma multiforme. Science 321: 1807–12.

[19] Andrewes C (1964) Tumour-viruses and virus-tumours. BMJ : 653–8.

[20] Davies H, Hunter C, Smith R, Stephens P, Greenman C, et al. (2005) Somatic mutations of the protein kinase gene family in human lung cancer. Cancer Res 65: 7591–5.

[21] Stephens P, Edkins S, Davies H, Greenman C, Cox C, et al. (2005) A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nat Genet 37: 590–2.

[22] Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF (2006) Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics 173: 2187–98.

[23] Malumbres M, Barbacid M (2001) To cycle or not to cycle: a critical decision in cancer. Nat Rev Cancer 1: 222–31.

[24] Malumbres M, Barbacid M (2007) Cell cycle kinases in cancer. Curr Opin Genet Dev 17: 60–5.

[25] Nguyen DX, Bos PD, Massagué J (2009) Metastasis: from dissemination to organ-specific colonization. Nat Rev Cancer 9: 274–84.

[26] Malumbres M, Barbacid M (2009) Cell cycle, cdks and cancer: a changing paradigm. Nat Rev Cancer 9: 153–66.

[27] Bamford S, Dawson E, Forbes S, Clements J, Pettett R, et al. (2004) The cosmic (catalogue of somatic mutations in cancer) database and website. Br J Cancer 91: 355–8.

[28] Schmidt-Kittler O, Ragg T, Daskalakis A, Granzow M, Ahr A, et al. (2003) From latent disseminated cells to overt metastasis: genetic analysis of systemic breast cancer progression. Proc Natl Acad Sci USA 100: 7737–42.

[29] Podsypanina K, Du YCN, Jechlinger M, Beverly LJ, Hambardzumyan D, et al. (2008) Seeding and propagation of untransformed mouse mammary cells in the lung. Science 321: 1841–4.

[30] Gray JW (2003) Evidence emerges for early metastasis and parallel evolution of primary and metastatic tumors. Cancer Cell 4: 4–6.

[31] Yokota J, Kohno T (2004) Molecular footprints of human lung cancer progression. Cancer Sci 95: 197–204.

[32] Bernards R, Weinberg RA (2002) A progression puzzle. Nature 418: 823.

[33] Nguyen DX, Massagué J (2007) Genetic determinants of cancer metastasis. Nat Rev Genet 8: 341–52.

[34] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. (2004) A census of human cancer genes. Nat Rev Cancer 4: 177–83.

[35] Garber K (2006) The second wave in kinase cancer drugs. Nat Biotechnol 24: 127–30.

[36] Noble MEM, Endicott JA, Johnson LN (2004) Protein kinase inhibitors: insights into drug design from structure. Science 303: 1800–5.

[37] Johnson L (2007) Protein kinases and their therapeutic exploitation. Biochem Soc Trans 35: 7–11.

[38] Fischer EH, Krebs EG (1955) Conversion of phosphorylase b to phosphorylase a in muscle extracts. J Biol Chem 216: 121–32.

[39] Wang JY (1998) Protein kinases entering the information age. J Biomed Sci 5: 73.

[40] Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298: 1912–34.

[41] Hunter T, Plowman GD (1997) The protein kinases of budding yeast: six score and more. Trends Biochem Sci 22: 18–22.

[42] Manning G (2005) Genomic overview of protein kinases. WormBook : 1–19.

[43] Manning G, Plowman GD, Hunter T, Sudarsanam S (2002) Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci 27: 514–20.

[44] Caenepeel S, Charydczak G, Sudarsanam S, Hunter T, Manning G (2004) The mouse kinome: discovery and comparative genomics of all mouse protein kinases. Proc Natl Acad Sci USA 101: 11707–12.

[45] Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. (2005) The universal protein resource (uniprot). Nucleic Acids Res 33: D154–9.

[46] Knight JDR, Qian B, Baker D, Kothary R (2007) Conservation, variability and the modeling of active protein kinases. PLoS ONE 2: e982.

[47] Kornev AP, Taylor SS, Eyck LFT (2008) A helix scaffold for the assembly of active protein kinases. Proc Natl Acad Sci USA 105: 14377–82.

[48] Russo AA, Jeffrey PD, Pavletich NP (1996) Structural basis of cyclin-dependent kinase activation by phosphorylation. Nat Struct Biol 3: 696–700.

[49] Yuan ZL, Guan YJ, Wang L, Wei W, Kane AB, et al. (2004) Central role of the threonine residue within the p+1 loop of receptor tyrosine kinase in stat3 constitutive phosphorylation in metastatic cancer cells. Mol Cell Biol 24: 9390–400.

[50] Kornev AP, Haste NM, Taylor SS, Eyck LFT (2006) Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. Proc Natl Acad Sci USA 103: 17783–8.

[51] Shi Z, Resing KA, Ahn NG (2006) Networks for the allosteric control of protein kinases. Curr Opin Struct Biol 16: 686–92.

[52] Hubbard SR (1997) Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and atp analog. EMBO J 16: 5572–81.

[53] Hubbard SR, Mohammadi M, Schlessinger J (1998) Autoregulatory mechanisms in protein-tyrosine kinases. J Biol Chem 273: 11987–90.

[54] Gonfloni S, Weijland A, Kretzschmar J, Superti-Furga G (2000) Crosstalk between the catalytic and regulatory domains allows bidirectional regulation of src. Nat Struct Biol 7: 281–6.

[55] Huse M, Kuriyan J (2002) The conformational plasticity of protein kinases. Cell 109: 275–82.

[56] Nagar B, Hantschel O, Young MA, Scheffzek K, Veach D, et al. (2003) Structural basis for the autoinhibition of c-abl tyrosine kinase. Cell 112: 859–71.

[57] Eyers PA, Erikson E, Chen LG, Maller JL (2003) A novel mechanism for activation of the protein kinase aurora a. Curr Biol 13: 691–7.

[58] Bishop JD, Schumacher JM (2002) Phosphorylation of the carboxyl terminus of inner centromere protein (incenp) by the aurora b kinase stimulates aurora b kinase activity. J Biol Chem 277: 27577–80.

[59] Zhang X, Gureasko J, Shen K, Cole PA, Kuriyan J (2006) An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. Cell 125: 1137–49.

[60] Dey M, Cao C, Dar AC, Tamura T, Ozato K, et al. (2005) Mechanistic link between pkr dimerization, autophosphorylation, and eif2alpha substrate recognition. Cell 122: 901–13.

[61] Dar AC, Dever TE, Sicheri F (2005) Higher-order substrate recognition of eif2alpha by the rna-dependent protein kinase pkr. Cell 122: 887–900.

[62] Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, et al. (2003) Targets of the cyclin-dependent kinase cdk1. Nature 425: 859–64.

[63] Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, et al. (2005) Global analysis of protein phosphorylation in yeast. Nature 438: 679–84.

[64] Jin J, Xie X, Chen C, Park JG, Stark C, et al. (2009) Eukaryotic protein domains as functional units of cellular evolution. Sci Signal 2: ra76.

[65] Peisajovich SG, Garbarino JE, Wei P, Lim WA (2010) Rapid diversification of cell signaling phenotypes by modular domain recombination. Science 328: 368–72.

[66] Apic G, Russell RB (2010) Domain recombination: a workhorse for evolutionary innovation. Sci Signal 3: pe30.

[67] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbsnp: the ncbi database of genetic variation. Nucleic Acids Res 29: 308–11.

[68] Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2011) Ensembl 2011. Nucleic Acids Res 39: D800–6.

[69] Consortium IH, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–8.

[70] Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) Mckusick's online mendelian inheritance in man (omim). Nucleic Acids Res 37: D793–6.

[71] Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, et al. (2004) The swiss-prot variant page and the modsnp database: a resource for sequence and structure information on human protein variants. Hum Mutat 23: 464–70.

[72] Bairoch A, Boeckmann B, Ferro S, Gasteiger E (2004) Swiss-prot: juggling between evolution and stability. Brief Bioinformatics 5: 39–55.

[73] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235–42.

[74] Hurst J, McMillan L, Porter C, Allen J, Fakorede A, et al. (2009) The saapdb web resource: A large-scale structural analysis of mutant proteins. Hum Mutat .

[75] Ortutay C, Väliaho J, Stenberg K, Vihinen M (2005) Kinmutbase: a registry of disease-causing mutations in protein kinase domains. Hum Mutat 25: 435–42.

[76] Richardson CJ, Gao Q, Mitsopoulous C, Zvelebil M, Pearl LH, et al. (2009) Mokca database–mutations of kinases in cancer. Nucleic Acids Res 37: D824–31.

[77] Blaschke C, Valencia A (2001) The potential use of suiseki as a protein interaction discovery tool. Genome Inform 12: 123–34.

[78] Krallinger M, Valencia A, Hirschman L (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol 9 Suppl 2: S8.

[79] Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A (2008) Overview of the protein-protein interaction annotation extraction task of biocreative ii. Genome Biol 9 Suppl 2: S4.

[80] Krallinger M, Leitner F, Valencia A (2010) Analysis of biological processes and diseases using text mining approaches. Methods Mol Biol 593: 341–82.

[81] Leitner F, Chatr-aryamontri A, Mardis SA, Ceol A, Krallinger M, et al. (2010) The febs letters/biocreative ii.5 experiment: making biological information accessible. Nat Biotechnol 28: 897–9.

[82] Rebholz-Schuhmann D, Marcel S, Albert S, Tolle R, Casari G, et al. (2004) Automatic extraction of mutations from medline and cross-validation with omim. Nucleic Acids Res 32: 135–42.

[83] Horn F, Lau AL, Cohen FE (2004) Automated extraction of mutation data from the literature: application of mutext to g protein-coupled receptors and nuclear hormone receptors. Bioinformatics 20: 557–68.

[84] Yip YL, Lachenal N, Pillet V, Veuthey AL (2007) Retrieving mutation-specific information for human proteins in uniprot/swiss-prot knowledgebase. J Bioinform Comput Biol 5: 1215–31.

[85] Lee LC, Horn F, Cohen FE (2007) Automatic extraction of protein point mutations using a graph bigram association. PLoS Comput Biol 3: e16.

[86] Baker CJO, Rene W (2006) Mutation mining - a prospector's tale. Journal of Information Systems Frontiers 8: 47–57.

[87] Erdogmus M, Sezerman OU (2007) Application of automatic mutation-gene pair extraction to diseases. J Bioinform Comput Biol 5: 1261–75.

[88] McDonald RT, Winters RS, Mandel M, Jin Y, White PS, et al. (2004) An entity tagger for recognizing acquired genomic variations in cancer literature. Bioinformatics 20: 3249–51.

[89] Furlong LI, Dach H, Hofmann-Apitius M, Sanz F (2008) Osirisv1.2: a named entity recognition system for sequence variants of genes in biomedical literature. BMC Bioinformatics 9: 84.

[90] Caporaso JG, Baumgartner WA, Randolph DA, Cohen KB, Hunter L (2007) Mutationfinder: a high-performance system for extracting point mutation mentions from text. Bioinformatics 23: 1862–5.

[91] Cline M, Karchin R (2010) Using bioinformatics to predict the functional impact of snvs. Bioinformatics .

[92] Karchin R (2009) Next generation tools for the annotation of human snps. Brief Bioinformatics 10: 35–52.

[93] Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11: 863–74.

[94] Clifford RJ, Edmonson MN, Nguyen C, Buetow KH (2004) Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics 20: 1006–14.

[95] Wang Z, Moult J (2001) Snps, protein structure, and disease. Hum Mutat 17: 263–70.

[96] Yue P, Moult J (2006) Identification and analysis of deleterious human snps. J Mol Biol 356: 1263–74.

[97] Yue P, Melamud E, Moult J (2006) Snps3d: candidate gene and snp selection for association studies. BMC Bioinformatics 7: 166.

[98] Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous snps: server and survey. Nucleic Acids Res 30: 3894–900.

[99] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248–9.

[100] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) Panther: a library of protein families and subfamilies indexed by function. Genome Res 13: 2129–41.

[101] Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, et al. (2005) Pmut: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21: 3176–8.

[102] Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, et al. (2005) Ls-snp: large-scale annotation of coding non-synonymous snps based on multiple information sources. Bioinformatics 21: 2814–20.

[103] Yue P, Li Z, Moult J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353: 459–73.

[104] Bromberg Y, Rost B (2007) Snap: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35: 3823–35.

[105] Kaminker JS, Zhang Y, Waugh A, Haverty PM, Peters B, et al. (2007) Distinguishing cancer-associated missense mutations from common polymorphisms. Cancer Res 67: 465–73.

[106] Kaminker JS, Zhang Y, Watanabe C, Zhang Z (2007) Canpredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res 35: W595–8.

[107] Torkamani A, Schork NJ (2007) Accurate prediction of deleterious protein kinase polymorphisms. Bioinformatics 23: 2918–25.

[108] Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30: 1237–44.

[109] Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res .

[110] Reva B, Antipin Y, Sander C (2007) Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol 8: R232.

[111] González-Pérez A, López-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous snvs with a consensus deleteriousness score, condel. Am J Hum Genet 88: 440–9.

[112] Stone EA, Sidow A (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 15: 978–86.

[113] Piirilä H, Väliaho J, Vihinen M (2006) Immunodeficiency mutation databases (idbases). Hum Mutat 27: 1200–8.

[114] Kwok CJ, Martin ACR, Au SWN, Lam VMS (2002) G6pddb, an integrated database of glucose-6-phosphate dehydrogenase (g6pd) mutations. Hum Mutat 19: 217–24.

[115] Kemball-Cook G, Tuddenham EG, Wacey AI (1998) The factor viii structure and mutation resource site: Hamsters version 4. Nucleic Acids Res 26: 216–9.

[116] Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, et al. (2007) Impact of mutant p53 functional properties on tp53 mutation patterns and tumor phenotype: lessons from recent developments in the iarc tp53 database. Hum Mutat 28: 622–9.

[117] Leigh SEA, Foster AH, Whittall RA, Hubbart CS, Humphries SE (2008) Update and analysis of the university college london low density lipoprotein receptor familial hypercholesterolemia database. Ann Hum Genet 72: 485–98.

[118] Tuchman M, Jaleel N, Morizono H, Sheehy L, Lynch MG (2002) Mutations and polymorphisms in the human ornithine transcarbamylase gene. Hum Mutat 19: 93–107.

[119] Wroe R, Butler AWL, Andersen PM, Powell JF, Al-Chalabi A (2008) Alsod: the amyotrophic lateral sclerosis online database. Amyotroph Lateral Scler 9: 249–50.

[120] Yeats C, Lees J, Reid A, Kellam P, Martin N, et al. (2008) Gene3d: comprehensive structural and functional annotation of genomes. Nucleic Acids Res 36: D414–8.

[121] Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, et al. (2007) The cath domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35: D291–7.

[122] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–402.

[123] Edgar RC (2004) Muscle: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.

[124] Prlic A, Down TA, Hubbard TJP (2005) Adding some spice to das. Bioinformatics 21 Suppl 2: ii40–1.

[125] Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal .

[126] Pei J, Grishin NV (2001) Al2co: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17: 700–12.

[127] López G, Valencia A, Tress ML (2007) Firedb–a database of functionally important residues from proteins of known structure. Nucleic Acids Res 35: D219–23.

[128] Porter CT, Bartlett GJ, Thornton JM (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32: D129–33.

[129] Casari G, Sander C, Valencia A (1995) A method to predict functional residues in proteins. Nat Struct Biol 2: 171–8.

[130] del Sol A, del Sol Mesa A, Pazos F, Valencia A (2003) Methods for predicting functionally important residues. J Mol Biol 326: 1289-1302.

[131] Rausell A, Juan D, Pazos F, Valencia A (2010) Protein interactions and ligand binding: from protein subfamilies to functional specificity. Proc Natl Acad Sci USA 107: 1995–2000.

[132] Greenacre MJ (1984) Theory and applications of correspondence analysis. Bell System Technical Journal .

[133] Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J Mol Biol 271: 511–23.

[134] Izarzugaza JMG, Graña O, Tress ML, Valencia A, Clarke ND (2007) Assessment of intramolecular contact predictions for casp7. Proteins 69 Suppl 8: 152–8.

[135] Ezkurdia I, Graña O, Izarzugaza J, Tress M (2009) Assessment of domain boundary predictions and the prediction of intramolecular contacts in casp8. Proteins .

[136] Care MA, Needham CJ, Bulpitt AJ, Westhead DR (2007) Deleterious snp prediction: be mindful of your training data! Bioinformatics 23: 664–72.

[137] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet 25: 25–9.

[138] Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The pfam protein families database. Nucleic Acids Res 38: D211–22.

[139] Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105–32.

[140] Ye ZQ, Zhao SQ, Gao G, Liu XQ, Langlois RE, et al. (2007) Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (sap). Bioinformatics 23: 1444–50.

[141] Wainreb G, Ashkenazy H, Bromberg Y, Starovolsky-Shitrit A, Haliloglu T, et al. (2010) Mud: an interactive web server for the prediction of non-neutral substitutions using protein structural data. Nucleic Acids Res 38 Suppl: W523–8.

[142] Diella F, Gould CM, Chica C, Via A, Gibson TJ (2008) Phospho.elm: a database of phosphorylation sites–update 2008. Nucleic Acids Res 36: D240–4.

[143] Izarzugaza JMG, Baresic A, McMillan LEM, Yeats C, Clegg AB, et al. (2009) An integrated approach to the interpretation of single amino acid polymorphisms within the framework of cath and gene3d. BMC Bioinformatics 10 Suppl 8: S5.

[144] Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, et al. (2006) Modbase: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 34: D291–5.

[145] Laskowski RA, Chistyakov VV, Thornton JM (2005) Pdbsum more: new summaries and analyses of the known 3d structures of proteins and nucleic acids. Nucleic Acids Res 33: D266–8.

[146] Loeys BL, Chen J, Neptune ER, Judge DP, Podowski M, et al. (2005) A syndrome of altered cardiovascular, craniofacial, neurocognitive and skeletal development caused by mutations in tgfbr1 or tgfbr2. Nat Genet 37: 275–81.

[147] Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, et al. (2002) Tumorigenesis: Raf/ras oncogenes and mismatch-repair status. Nature 418: 934.

[148] Davies H, Bignell GR, Cox C, Stephens P, Edkins S, et al. (2002) Mutations of the braf gene in human cancer. Nature 417: 949–54.

[149] Lee JW, Yoo NJ, Soung YH, Kim HS, Park WS, et al. (2003) Braf mutations in non-hodgkin's lymphoma. Br J Cancer 89: 1958–60.

[150] Izarzugaza JMG, Krallinger M, Rodriguez-Penagos C, Valencia A (2009) Extraction of human kinase mutations from literature, databases and genotyping studies. BMC Bioinformatics 10 Suppl 8: S1.

[151] Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, et al. (2008) Overview of biocreative ii gene normalization. Genome Biol 9 Suppl 2: S3.

[152] Tam IYS, Chung LP, Suen WS, Wang E, Wong MCM, et al. (2006) Distinct epidermal growth factor receptor and kras mutation patterns in non-small cell lung cancer patients with different tobacco exposure and clinicopathologic features. Clin Cancer Res 12: 1647–53.

[153] Graña O, Baker D, MacCallum RM, Meiler J, Punta M, et al. (2005) Casp6 assessment of contact prediction. Proteins 61 Suppl 7: 214–24.

[154] Izarzugaza J, Redfern O, Orengo C, Valencia A (2009) Cancer-associated mutations are preferentially distributed in protein kinase functional sites. Proteins .

[155] Izarzugaza JMG, McMillan LEM, Baresic A, Orengo CA, Martin ACR, et al. (2011) Characterization of pathogenic germline mutations in human protein kinases. BMC Bioinformatics 12 Suppl 4.

[156] Fiser A, Sali A (2003) Modeller: generation and refinement of homology-based protein structure models. Meth Enzymol 374: 461–91.

[157] Pazos F, Rausell A, Valencia A (2006) Phylogeny-independent detection of functional residues. Bioinformatics 22: 1440–8.

[158] Dhillon AS, Hagan S, Rath O, Kolch W (2007) Map kinase signalling pathways in cancer. Oncogene 26: 3279–90.

[159] Shannon CE (1997) The mathematical theory of communication. 1963. MD Comput 14: 306–17.

[160] Caceres M, Teran CG, Rodriguez S, Medina M (2008) Prevalence of insulin resistance and its association with metabolic syndrome criteria among bolivian children and adolescents with obesity. BMC Pediatr 8: 31.

[161] Leroy JG, Nuytinck L, Lambert J, Naeyaert JM, Mortier GR (2007) Acanthosis nigricans in a child with mild osteochondrodysplasia and k650q mutation in the fgfr3 gene. Am J Med Genet A 143A: 3144–9.

[162] Zankl A, Elakis G, Susman RD, Inglis G, Gardener G, et al. (2008) Prenatal and postnatal presentation of severe achondroplasia with developmental delay and acanthosis nigricans (saddan) due to the fgfr3 lys650met mutation. Am J Med Genet A 146A: 212–8.

[163] Fonseca R, Costa-Lima MA, Cosentino V, Orioli IM (2008) Second case of beare-stevenson syndrome with an fgfr2 ser372cys mutation. Am J Med Genet A 146A: 658–60.

[164] Yue P, Melamud E, Moult J (2006) Snps3d: candidate gene and snp selection for association studies. BMC Bioinformatics 7: 166.

[165] Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Support Vector Machines 46: 389–422.

[166] Rakotomamonjy A (2003) Variable selection using svm-based criteria. JMLR 3: 1357–70.

[167] Uzun A, Leslin CM, Abyzov A, Ilyin V (2007) Structure snp (stsnp): a web server for mapping and modeling nssnps on protein structures with linkage to metabolic pathways. Nucleic Acids Res 35: W384–92.

[168] Krallinger M, Vazquez M, Leitner F, Salgado D, Chatraryamontri A, et al. (2011) The protein-protein interaction tasks of biocreative iii: Classification/ranking of articles and linking bio-ontology concepts to full text. BMC Bioinformatics : in press.

[169] Kinnings SL, Jackson RM (2009) Binding site similarity analysis for the functional classification of the protein kinase family. Journal of chemical information and modeling .

[170] Taylor SS, Kornev AP (2011) Protein kinases: evolution of dynamic regulatory proteins. Trends Biochem Sci 36: 65–77.

[171] Lahiry P, Torkamani A, Schork NJ, Hegele RA (2010) Kinase mutations in human disease: interpreting genotype-phenotype relationships. Nat Rev Genet 11: 60–74.

[172] Torkamani A, Kannan N, Taylor SS, Schork NJ (2008) Congenital disease snps target lineage specific structural elements in protein kinases. Proc Natl Acad Sci USA 105: 9011–6.

[173] Tartaglia M, Mehler EL, Goldberg R, Zampino G, Brunner HG, et al. (2001) Mutations in ptpn11, encoding the protein tyrosine phosphatase shp-2, cause noonan syndrome. Nat Genet 29: 465–8.

[174] Wan PTC, Garnett MJ, Roe SM, Lee S, Niculescu-Duvaz D, et al. (2004) Mechanism of activation of the raf-erk signaling pathway by oncogenic mutations of b-raf. Cell 116: 855–67.

[175] Rodriguez-Viciana P, Tetsu O, Tidyman WE, Estep AL, Conger BA, et al. (2006) Germline mutations in genes within the mapk pathway cause cardio-facio-cutaneous syndrome. Science 311: 1287–90.

[176] Plaza-Menacho I, Burzynski GM, de Groot JW, Eggen BJL, Hofstra RMW (2006) Current concepts in ret-related genetics, signaling and therapeutics. Trends Genet 22: 627–36.

[177] Romeo G, Ronchetto P, Luo Y, Barone V, Seri M, et al. (1994) Point mutations affecting the tyrosine kinase domain of the ret proto-oncogene in hirschsprung's disease. Nature 367: 377–8.

[178] Ferrer-Costa C, Orozco M, de la Cruz X (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 315: 771–86.

[179] Capriotti E, Fariselli P, Casadio R (2005) I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33: W306–10.

[180] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33: D514–7.

[181] Cooper DN, Stenson PD, Chuzhanova NA (2006) The human gene mutation database (hgmd) and its exploitation in the study of mutational mechanisms. Curr Protoc Bioinformatics Chapter 1: Unit 1.13.

[182] Fredman D, Siegfried M, Yuan YP, Bork P, Lehväslaiho H, et al. (2002) Hgvbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. Nucleic Acids Res 30: 387–91.

[183] Sunyaev S, Ramensky V, Koch I, Lathe W, Kondrashov AS, et al. (2001) Prediction of deleterious human alleles. Hum Mol Genet 10: 591–7.

[184] Chasman D, Adams RM (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307: 683–706.

[185] Orengo C (1999) Cora–topological fingerprints for protein structural families. Protein Sci 8: 699–715.

# Index of Figures

# Index of Tables

# CURRICULUM VITAE

*José María González-Izarzugaza Martínez*

## EMPLOYMENT HISTORY

**2006 – Currently** **CNIO – Spanish National Cancer Research Centre**
Structural Computational Biology Group
Prof. Alfonso Valencia Herrera

**2008** **UCL – University College London**
Biomolecular Structure and Modelling Unit
Prof. Christine A. Orengo

**2005 – 2006** **CNB-CSIC – Spanish National Centre for Biotechnology**
Protein Design Group
Prof. Alfonso Valencia Herrera

**2004 – 2005** **Noray Bioinformatics SL**
Genomics and Proteomics Group
Dr. Julio Font Pérez

**2002 – 2003** **Agilent Technologies Ltd.**
Life Sciences Business Unit
Mr. José María Molina Tejada

## EDUCATION

**2005 – Currently** **Predoctoral Fellowship in Molecular Biology**
Universidad Autónoma de Madrid

**2004** **Master of Science in Bioinformatics and Computational Biology**
Universidad Complutense de Madrid

**2000 – 2003** **Bachelor of Science in Biochemistry**
Universidad del País Vasco

**1997 – 2000** **Degree on Chemical Sciences**
Universidad del País Vasco

**http://jmgi.tk**

**PUBLICATIONS**

- **Izarzugaza JMG**, Pozo A, Vazquez M, Valencia A. Prioritization of pathogenic mutations in the Protein Kinase superfamily. *In preparation*.
- **Izarzugaza JMG**, Baresic A, McMillan LEM, Orengo CA, Martin ACR, Valencia A. Characterization of pathogenic germline mutations inhuman Protein Kinases. BMC Bioinformatics 2011, 12:336
- **Izarzugaza JMG**, Baresic A, McMillan LEM, Yeats C, Clegg AB, Orengo CA, Martin ACR, Valencia A. An integrated approach to the intepretation of Single Amino Acid Polymorphisms within the framework of CATH and Gene3D. BMC Bioinformatics 2009, 10(Suppl 8):I1
- **Izarzugaza JMG**, Krallinger M, Rodriguez-Penagos C, Valencia A. Extraction of human kinase mutations from literature, databases and genotyping studies. BMC Bioinformatics 2009, 10(Suppl 8):II
- Ezkurdia I, Graña O, **Izarzugaza JM**, Tress ML. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. Proteins 2009 Jul 24.
- **Izarzugaza JM**, Redfern OC, Orengo CA, Valencia A. Cancer associated mutations are preferentially distributed in protein kinase functional sites. Proteins 2009 Jun 19
- Baudot A, Real F, **Izarzugaza J**, Valencia A. From cancer genomes to cancer models: bridging the gaps. EMBO Rep (2009) pp.
- Pazos F, Juan D, **Izarzugaza JM**, Leon E, Valencia A. Prediction of protein interaction based on similarity of phylogenetic trees. Methods Mol Biol (2008) vol. 484 pp. 523-35
- **Izarzugaza JM**, Juan D, Pons C, Pazos F, Valencia A. Enhancing the prediction of protein pairing between interacting families using orthology information. BMC Bioinformatics. 2008 9:35
- **Izarzugaza JM**, Graña O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7 Proteins 2007 Aug 1
- **Izarzugaza JM**, Juan D, Pons C, Ranea JA, Valencia A, Pazos F. TSEMA: interactive prediction of protein pairings between interacting families. Nucleic Acids Res. 2006 Jul 1;34 (Web Servers issue):W315-9

**ORAL COMMUNICATIONS**

- Prediction of the pathogenicity of mutations within the human kinome
  Cancer-om-atics II. Madrid (Spain), 2011
- Characterization of pathogenic germline mutations in human protein kinases.
  AIMM/ECCB. Ghent (Belgium), 2010
- Cancer-associated mutations are preferentially distributed in protein kinase functional sites.
  University College London, London (UK), 2008
- TSEMA The Server for Effective Mapping Assessment.
  School of Bioinformatics. Universitá di Bologna. Bologna (Italy). 2006
- TSEMA The Server for Effective Mapping Assessment.
  SAC-CASP7. Asilomar (California, USA), 2006
- TSEMA The Server for Effective Mapping Assessment.
  Spanish National Thematic Network for Structure and Folding. Bilbao (Spain), 2006

**POSTERS**

- **Izarzugaza JM**, Pozo A, Vazquez M, Valencia A. KinMut: Predicting the pathogenicity of mutations within the human kinome. ECCB. 2011.
- Vazquez M, **Izarzugaza JM**, Torre V, Valencia A. Bioinformatics for personalized cancer treatment. ECCB, 2011.
- **Izarzugaza JM**, Pozo A, Vazquez M, Valencia A. Prediction of the pathogenicity of mutations within the human kinome. Cancer-om-atics II. 2011.
- Vazquez M, **Izarzugaza JM**, Torre V, Valencia A. Bioinformatics for personalized cancer treatment. Cancer-om-atics II, 2011.
- Baresic A, Alnumair N, **Izarzugaza JM**, Valencia A, Martin ACR. SAAPdb: Structural Analysis of single amino acid polymorphisms. AIMM/ECCB, 2010
- **Izarzugaza JM**, Juan D, Pons C, Pazos F, Valencia A. TAG-TSEMA The Server for Effective Mapping Assessment. ECCB, 2008.
- **Izarzugaza JM**, Juan D, Pons C, Ranea JA, Valencia A, Pazos F. TSEMA The Server for Effective Mapping Assessment, CASP7, 2006.

**TEACHING EXPERIENCE**

- Masters Degree in Molecular and Cellular Biology, Universidad Autonoma de Madrid. Advanced Bioinformatics and Systems Biology, 2009 - 2011
- Masters Degree in Bioinformatics and Computational Biology, Universidad Complutense de Madrid. Sequence Analysis and Databases, 2007 - 2011
- Postgraduate courses in Molecular Biology. Universidad Autonoma de Madrid. Sequence Analysis and Databases, 2008 - 2009
- Summer school of Bioinformatics, Universidad Complutense de Madrid. Sequence Analysis and Databases, 2006 - 2007

**OTHER MERITS AND CONTRIBUTIONS**

- CASP Residue-Residue Contact Prediction Assessor, 2006 - 2008
- Expert on Biotechnology and Pharmaceutical Industry for the implementation of the Technological Development web-portal for the Council of Madrid (MadriTec) 2006
- Referee for several Scientific Journals and Congresses, 2004 - Current

**LANGUAGE SKILLS**

- Spanish Native Speaker
- English Fluent; Cambridge Advanced Certificate (2003)
- German Elementary
- Basque Intermediate

BMC Bioinformatics

BioMed Central

Research

Open Access

# An integrated approach to the interpretation of Single Amino Acid Polymorphisms within the framework of CATH and Gene3D

Jose MG Izarzugaza*[1,2], Anja Baresic[1], Lisa EM McMillan[1], Corin Yeats[1], Andrew B Clegg[1], Christine A Orengo[1], Andrew CR Martin[1] and Alfonso Valencia[2]

Address: [1]Institute of Structural and Molecular Biology, Darwin Building, University College London, Gower Street, London WC1E 6BT, UK and [2]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernandez Almagro 3, E28029 Madrid, Spain

Email: Jose MG Izarzugaza* - jmgonzalez@cnio.es; Anja Baresic - anya@biochem.ucl.ac.uk; Lisa EM McMillan - mcmillan@biochem.ucl.ac.uk; Corin Yeats - yeats@biochem.ucl.ac.uk; Andrew B Clegg - clegg@biochem.ucl.ac.uk; Christine A Orengo - orengo@biochem.ucl.ac.uk; Andrew CR Martin - a.martin@biochem.ucl.ac.uk; Alfonso Valencia - avalencia@cnio.es

* Corresponding author

*from* ECCB 2008 Workshop: Annotations, interpretation and management of mutations (AIMM)
Cagliari, Italy. 22 September 2008

Published: 27 August 2009

*BMC Bioinformatics* 2009, **10**(Suppl 8):S5 doi:10.1186/1471-2105-10-S8-S5

This article is available from: http://www.biomedcentral.com/1471-2105/10/S8/S5



## Abstract

**Background:** The phenotypic effects of sequence variations in protein-coding regions come about primarily via their effects on the resulting structures, for example by disrupting active sites or affecting structural stability. In order better to understand the mechanisms behind known mutant phenotypes, and predict the effects of novel variations, biologists need tools to gauge the impacts of DNA mutations in terms of their structural manifestation. Although many mutations occur within domains whose structure has been solved, many more occur within genes whose protein products have not been structurally characterized.

**Results:** Here we present 3DSim (3D Structural Implication of Mutations), a database and web application facilitating the localization and visualization of single amino acid polymorphisms (SAAPs) mapped to protein structures even where the structure of the protein of interest is unknown. The server displays information on 6514 point mutations, 4865 of them known to be associated with disease. These polymorphisms are drawn from SAAPdb, which aggregates data from various sources including dbSNP and several pathogenic mutation databases. While the SAAPdb interface displays mutations on known structures, 3DSim projects mutations onto known sequence domains in Gene3D. This resource contains sequences annotated with domains predicted to belong to structural families in the CATH database. Mappings between domain sequences in Gene3D and known structures in CATH are obtained using a MUSCLE alignment. 1210 three-dimensional structures corresponding to CATH structural domains are currently included in 3DSim; these domains are distributed across 396 CATH superfamilies, and provide a comprehensive overview of the distribution of mutations in structural space.

**Conclusion:** The server is publicly available at http://3DSim.bioinfo.cnio.es/. In addition, the database containing the mapping between SAAPdb, Gene3D and CATH is available on request and most of the functionality is available through programmatic web service access.

## Background

The most common biologically-relevant mutations are single base changes often referred to as **single nucleotide polymorphisms** (SNPs). These account for about 90% of sequence polymorphisms in humans [1] at an overall frequency of about one per 1000 bases [2]. Traditionally, SNPs are classified as coding or non-coding according to their genomic location – coding SNPs are further sub-classified according to the protein product expressed. Non-Synonymous SNPs (nsSNPs) are those that alter the amino acid sequence of the protein product, either through amino acid substitution (a 'single amino acid polymorphisms', SAAP), or by the generation of truncation mutations. By contrast, synonymous SNPs (also referred to as silent or sSNPs) are those that do not alter the amino acid sequence of the protein product.

Not all synonymous SNPs are neutral since they may still affect the expression of gene products or protein translation by introducing alterations into regulatory regions, interfering with splice sites or impinging on other regulatory mechanisms [3,4]. Similarly, not all nsSNPs are associated with pathological diseases, since some changes are, by nature, milder than others, and diseases commonly involve complex sets of alterations.

Strictly the term 'SNP' is defined as a mutation which occurs in at least 1% of a 'normal' population. Thus SNPs are expected to have a neutral non-deleterious or low-penetrance phenotypic effect whereas the term **pathogenic deviation** (PD) refers to those mutations that generally occur at much lower frequencies in the population and have a severe effect on phenotype.

The most commonly used database for storing information on SNPs is dbSNP [5], which currently contains several million validated SNPs from humans and other species. Other sources of genomic information about SNPs include Ensembl [6] and the HapMap Project [7].

Several efforts have been devoted to the prediction of the pathogenicity of amino acid mutations, resulting from single nucleotide changes. These methods make use of a set of characteristics which may be based both on sequence and structure, to determine whether a mutation can affect protein function and therefore be, potentially, associated with disease. This is an area of active research as shown by the considerable number of publications on the subject during the last few years [8-18].

Several efforts, SAAPdb [19] among others, have been devoted to compiling this information and to providing a sequence and structural analysis, where possible, aiming to determine the origin of the pathogenicity shown. In this type of repository, the term SNP is used to refer to essentially phenotypically silent mutations, while PD is used for mutations known to have a severe effect on phenotype, i.e. any single base change reported to correlate with disease. Online Mendelian Inheritance in Man (OMIM) [20] is a collection of information about inherited disease and contains data on PDs. However a great deal more information is held and maintained by individual research groups in locus-specific mutation databases or LSMDBs [21]. Like PDs, nsSNPs are point mutations, but by definition they occur in at least 1% of a 'normal' population. They are expected to have a neutral non-deleterious or low-pentrance phenotypic effect whereas PDs are known to be detrimental. By mapping these SAAPs (a term we use for both PDs and mutations resulting from nsSNPs) onto protein structures, we can begin to understand how protein structure might be affected by mutant residues, and so begin to explain the functional effect (if any) of the mutation. SAAPdb provides potential explanations for both PDs, derived from various sources, and SNPs, derived from dbSNP [5].

The CATH [22] structural domain database is a manually curated classification of domain structures found in the Protein Data Bank (PDB) [23], grouped according to evolutionary relationships and structural features. Hidden Markov Models (HMMs) are derived from alignments of these structural exemplars and used by Gene3D [24] to identify homologues within the protein sequences of UniProt [25], RefSeq [20] and Ensembl [6].

Here we present 3DSim (3D Structural Implication of Mutations), a system mapping single amino-acid polymorphisms onto structures of CATH domains. For sequences with no known structure, the Gene3D resource of domain structure annotations is used to map the sequence onto the closest homologous domain of known structure in CATH. Thus 3DSim is of particular interest when no structural information is available for a protein in which mutations are known to occur as it uses sequence homology to map to the closest representative structure. This provides a comprehensive overview of the distribution of mutations in structural space, as well as a visualization tool for pinpointing the locations of mutations on individual structures rendered in Jmol http://www.jmol.org/, as well as links to detailed information on each sequence, structure and mutation. The 3DSim application, which was designed with the aim of being very intuitive, easy to use and user-friendly, is publicly available at http://3DSim.bioinfo.cnio.es/. Several worked examples are available, along with a 6-minute video tutorial. In addition, for those advanced users needing intensive programmatic access to the information stored, the underlying database containing the mappings between SAAPdb, Gene3D and CATH is available on

request, and most of the functionality is available as web services implemented in SOAP.

## Results and discussion

### *The mapping between SAAPdb and Gene3D*

SAAPdb contains polymorphism data for 11956 sequences without a structure. Almost all of these could be mapped to Gene3D: 11904 identical sequences were found in the Gene3D database. Of the remaining 52, 17 had sequences with the same length and associated uniprot accession, leaving only 35 for which a reliable match could not be obtained directly.

### *The mapping between Gene3D and CATH*

Where no structural data are available, the best representative CATH domain is selected on the basis of homology. For each of the 2179 superfamilies in CATH, a database of all CATH domains was built. For each of the 11904 Gene3D domain sequences mapped to CATH structural superfamilies for which there is information about mutations in SAAPdb (see previous section), a BLAST [26] search was run against the corresponding superfamily database. The closest relative found (i.e. the one with the lowest e-value and highest sequence identity) was used to cluster the sequences. Sequences with a sequence identity less than 20% were placed in separate clusters. This process yielded 2091 different groups. The groups (including the sequence of the representative structure) were then aligned using MUSCLE [27] and the resulting alignments used to transfer the mutations from Gene3D sequences to CATH domain representative structures. At the end of the pipeline we were able to display information on 6514 point mutations, 4865 of them known to be associated with disease, mapping to 396 CATH superfamilies. The complete pipeline is described in Figure 1 and details are provided in the Methods section.

### *Description of web application*

The initial input for the system is a CATH superfamily identifier for which the user wants to retrieve information on mapped mutations. Alternatively, the database can be searched using Uniprot accession codes or CATH domain identifiers. The user can either manually introduce the desired superfamily identifier in the provided form, or browse the superfamilies in CATH in order to access the information. After this initial step, information about the

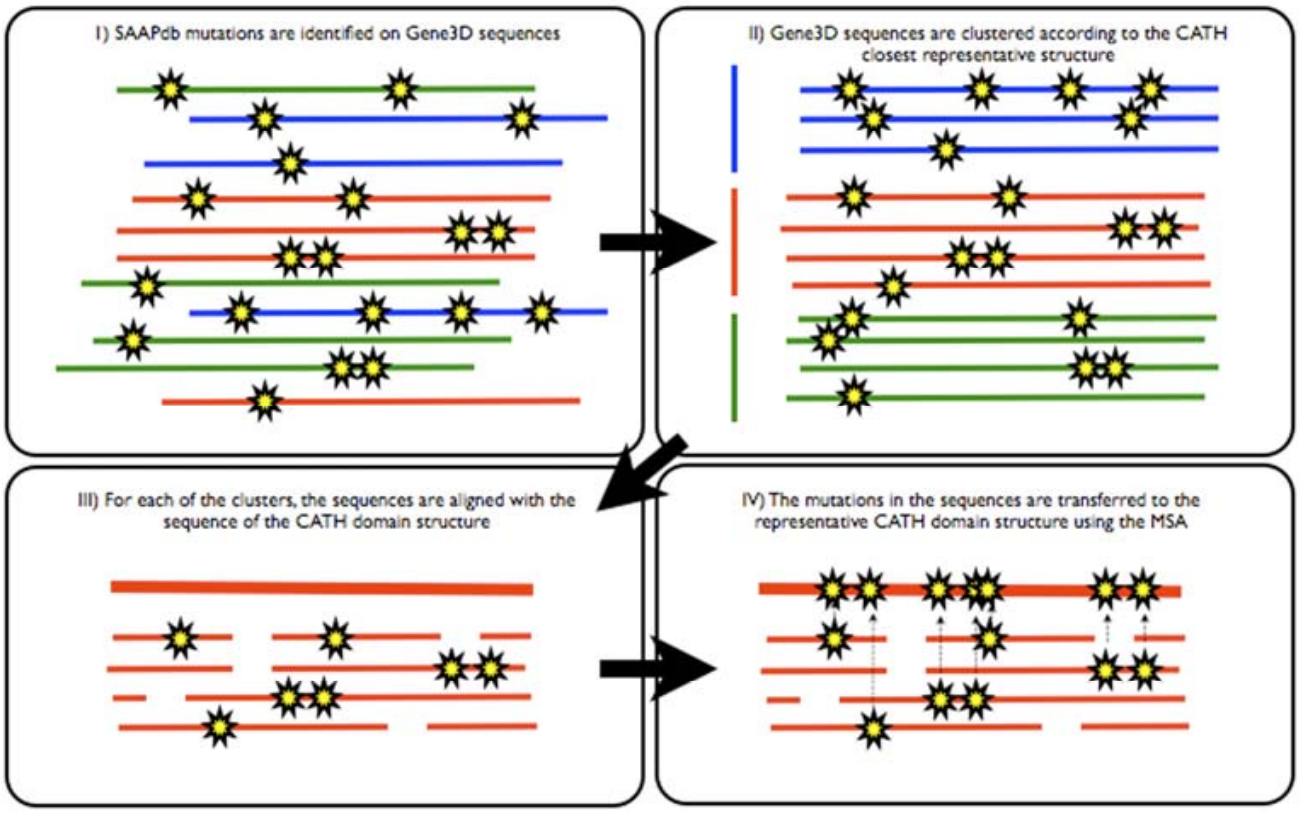



**Figure 1**
**The mapping pipeline**. Illustration of the mapping process between SAAPdb and CATH via Gene3D

selected CATH superfamily is displayed along with the CATH domains for which there is information about mutations in SAAPdb.

In addition, for users that are interested in a general overview of the distribution of mutations within structural superfamilies in CATH, one can obtain a list of the superfamilies with known mutations and analyze domains in that superfamily.

Once the user has selected a CATH domain, 3DSim displays both an interactive Jmol plug-in that allows the visualization of the mutations projected onto the three-dimensional structure of the representative CATH domain and a table displaying all the information available for that given domain in terms of available mutations, sequence and structure positions of the mutations, pathogenicity information, and similarity (BLAST sequence identity) between the sequences in Gene3D and the representative CATH domain sequence.

This similarity index provides the user with a hint about the reliability of the homology based transference of mutations between sequences in Gene3D and the structures in CATH. As a rule of thumb, the higher the similarity the more reliable the transference of mutations is. Tweaking this index is of particular interest when there are few mutations in the close relatives for a given structural family and looser constrains need to be taken into account to allow more mutations in the analysis. By default, the server rejects those mutations transferred from sequences obtaining a BLAST sequence identity of less than 20%, but – due to the interactive approach of the server – the user can decide to establish more stringent constrains depending on the study case.

In addition, the site is linked to several external annotation providers (including CATH, Gene3D, SAAPdb, Modbase, PDBsum and UniProt) where more information about the mutations, the proteins and the structures can be gathered. In particular, SAAPdb provides information about the structural implications of mutations. This information can be related, in some cases, to the pathogenic character of the mutations and provides an insight into the mechanism of molecular function for several proteins.

Figure 2 shows a worked example of the different views available through the server's graphical user interface.

### *Description of web services*

In order to allow remote programmatic access to the information contained in the database, we have developed a total of nine SOAP web services, powered by the Perl SOAP::Lite toolkit http://www.soaplite.com/. These allow users to retrieve in simple XML format:

- all known mutations for a given CATH domain, grouped by UniProt ID.
- the total number of mutations in a CATH domain.
- all the CATH domains which are associated with a given UniProt ID.
- the amino-acid sequence of a given CATH domain.
- all CATH domains in a CATH superfamily, queried by the four-part CATH code.
- the superfamily to which a given CATH domain belongs.
- the description and representative structure associated with a given CATH superfamily.
- all the mutations in SAAPdb for a given UniProt accession.
- the total number of mutations in SAAPdb for a given UniProt accession.

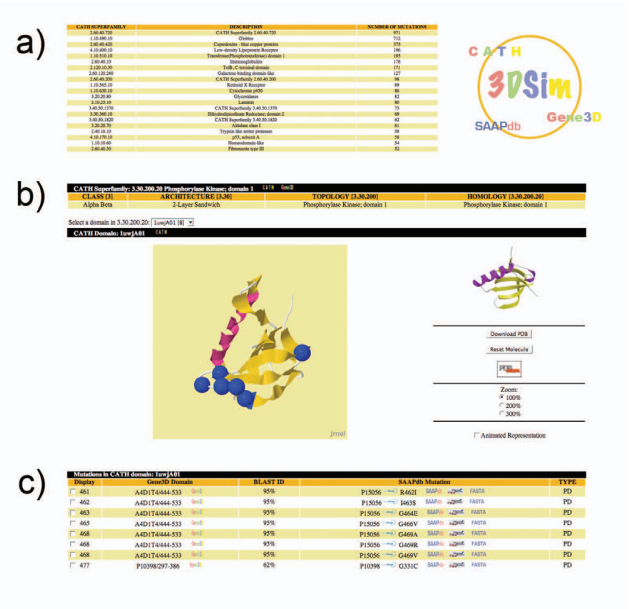



**Figure 2**
**The web interface**. A collage showing a worked example of the different sections available through the server's graphical user interface. a) Browsable list of superfamilies b) Example of structure displaying known pathogenic mutations c) Explanatory table of the mutations

These services were designed in such a way as to facilitate construction of computational analysis pipelines. For example, a user starting with a UniProt protein of interest could retrieve a list of all the domains found in that protein, then the CATH superfamilies to which each domain belongs, then all the other domains found in each superfamily, and filly all known mutations in those related domains, by chaining together four web service calls.

More information is provided at http://3DSim.bioinfo.cnio.es/webservices.html. The page contains example Perl code for querying the web-services, and examples of output from each one.

### *Database update*

The database storing the information presented by both the webserver and the webservice intrinsically depends on the other databases providing the source information (i.e. CATH, Gene3D and SAAPdb), each one being updated at its own pace. This fact, in addition to the computationally expensive calculations needed to compute the mapping between Gene3D and the representative structures in CATH, makes it impossible to schedule an automatic updating calendar. Therefore, the database will be updated based on a release system, where new versions will be made public as regularly as possible.

### *Typical usage example*

As an illustrative example, here we present the case of the ATP binding subunit of the kinases (CATH superfamily 1.10.510.10) which is accessible through the server's web page http://3DSim.bioinfo.cnio.es. This superfamily corresponds to the Phosphotransferase domain I homology group in CATH, and is subdivided into a number of different domains. However, for this particular example, we will focus only on the domain with the highest number of mutations (24), 1rw8A02. Of these 24 mutations, only three come from the sequence which maps directly to the domain. The remaining 21 come from homologous sequences with 40–65% sequence identity identified via Gene3D (Table 1). Figure 3 shows the structure with the pathogenic deviations coloured in blue. This image can be obtained directly from the server, and is one of the main features available for the analysis of the distribution of mutations within structures. Additional links to other structure-based databases such as PDBsum [28] are provided in order to enhance the information provided, for this particular case, the position of the catalytic site, involved in binding of ATP, is described to be near residues from 333 to 338. Visual inspection of the position of the pathogenic deviations reveals that they tend to cluster around this catalytic core of the structure. Indeed, the higher the similarity in terms of BLAST identity between the CATH domain and the Gene3D sequence, the closer these positions are to the binding core and hence, more reliable the observations are.

This PDB chain (1rw8A) maps to the UniProtKB/Swiss-Prot accession P36897 and the information provided by the UniProt record (accessible through the web server's cross references) shows that it corresponds to the TGF-beta receptor type-1 precursor in humans (TGFR1_HUMAN) for which there is a level of association with disease, in particular to Furlong syndrome also known as Loeys-Dietz syndrome type 1A (LDS1); [29]. LDS1 is an aortic aneurysm syndrome with widespread systemic involvement. The disorder is characterized by arterial tortuosity and aneurysms, craniosynostosis, hypertelorism, and bifid uvula (cleft palate). Other findings include exotropy, micrognathia and retrognathia, structural brain abnormalities, intellectual deficit, congenital heart disease, translucent skin, joint hyperlaxity and aneurysm with dissection throughout the arterial tree. The mutations listed as pathogenic deviations (R487P, M318R and D400G) in the server for this protein, which has a 100% identity between the Gene3D sequence and the representative structure of the CATH domain are reported in the literature [29] as involved in LDS1 development.

**Table 1: Mutations mapped to 1rw8A02. SwissProt accession P36897 maps directly to PDB code 1rw8 chain A and to CATH domain 1rw8A02 which represents residues 285–500. Other SwissProt entries containing known pathogenic deviations (PDs) are mapped to this domain via Gene3D and the mutations are mapped to the 1rw8 structure.**

| SwissProt Accession | Domain Range | Sequence Identity | Mutations |
|---|---|---|---|
| P36897 | 285–500 | 100% | M318R, D400G, R487P |
| O00238 | 284–499 | 65% | R486W |
| P36894 | 314–529 | 62% | A338D, C376Y, M470T |
| P37023 | 282–497 | 60% | C344Y, R374W, M376R<br>I398N, W399S, R411P<br>R411Q, R411W, R484W |
| P37173 | 330–546 | 40% | Y336N, A355P, G357W<br>S449F, E526Q, R528C<br>R528H, R537C |

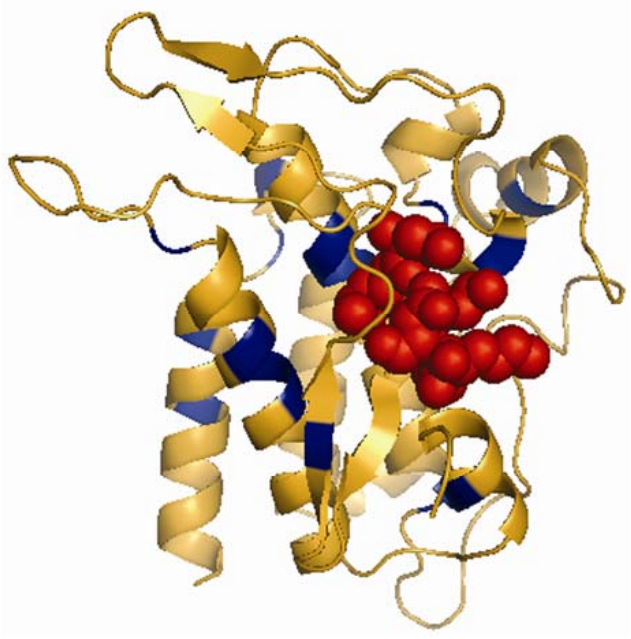


**Figure 3**
**Three-dimensional structure of CATH domain** 1rw8A02. 1rw8A02 is member of the Phosphotransferase superfamily in CATH (1.10.510.10). 24 positions are reported to be pathogenic mutations (PDs, blue).

## Conclusion

We have presented 3DSim (3D Structural Implication of Mutations), a system that enables the localization and visualization of single amino acid polymorphisms projected onto protein structures based on homologous relationships captured in the CATH and Gene3D databases. This provides a comprehensive overview of the distribution of mutations in structural space.

Although there are other servers mapping mutations to structure already (reviewed by Uzun *et al.* [30]) the server presented here has several unique features not available in existing servers. Firstly, the similar treatment of SNPs and the rarer more harmful PDs allows users to inspect and compare both kinds of mutation through the same interface, including explanatory metadata where available. Secondly, the localization of these SAAPs within the CATH hierarchy allows users to query and explore the distribution of mutations at various levels of structural classification. Thirdly, the mapping of sequences onto homologous CATH domains via Gene3D helps users predict the effects of polymorphisms in proteins whose structure has not been solved. Finally, the availability of the data via web services and database dumps enables power users to include this information efficiently in their own analyses. These facilities allow the independent integration of our data in any other pipeline or workflow.

The server has been running internally since we started working on the analysis of point mutations in protein families [31,32] and is accessible at http://3DSim.bioinfo.cnio.es/. Examples and documentation are also available, together with a tutorial video and samples of outputs of the main functions. This website is available to all users with no login requirement. It is likely that we will include additional features related with the structural interpretation of mutations and their relationship with disease, after receiving feedback from external users.

In summary the 3DSim server provides up-to-date, complete information automatically to map mutations in the domain sequences of proteins annotated in Gene3D onto protein structures classified in the CATH database.

## Methods

### *Obtaining the mutations from SAAPdb*

SAAPdb [19] is a database of single amino acid polymorphisms (SAAPs) from several resources, such as dbSNP [5], ADABase [33], G6PD [34] HAMSTeRs [35], IARC p53 Database [36], LDLR [37], OMIM http://www.ncbi.nlm.nih.gov/omim/, OTC [38], SOD1db [39] and ZAP70Base [33], mapped to protein structure, where available in the PDB [23]. As of October 2008, SAAPdb contains 9060 unique pathogenic deviations (PDs: SAAPs associated to a disease) and 2532 unique single nucleotide polymorphisms (SNPs: SAAPs with no known pathogenic effect) successfully mapped to the UniProtKB [25] sequences in Gene3D [24].

Both pathogenic deviations and single nucleotide polymorphisms were only taken into account if the alteration introduced was non-silent, that is, if the mutation is both in a coding residue and the resulting amino acid is different from the native one. Where mutations are recorded both as neutral and disease-associated, the mutations were considered pathogenic.

### *Gene3D domain assignments*

The process by which homologues of CATH domains are identified in sequences, and presented in the Gene3D database, has been described previously [40]. For this particular dataset, the CATH v3.2.0 Hmmer HMM library was scanned against the UniProt (Swiss-Prot and TrEMBL) sequence database in collaboration with the SIMAP database [41]. FASTA files of each superfamily were generated by extracting the subsequences of the domains belonging to each superfamily.

### *Gathering the sequences from Gene3D*

For each CATH superfamily a library of one or more HMMs is generated using the SAM Target2K procedure [22]. These HMMs are then searched against UniProt in collaboration with the SIMAP resource at the Munich

Information Centre for Protein Sequences [41]. The hits are resolved into a single set of non-overlapping domains for each sequence, using the in-house DomainFinder 2.0 protocol. The resulting domain subsequences were then extracted and dumped into the relevant superfamily FASTA file.

#### *Gathering the sequences of the CATH domains*

For each of the 2097 superfamiles in CATH [22], all CATH domains were recovered along with the corresponding amino acid sequences directly from CATH's Oracle database. A total of 86463 CATH domains were found. Afterwards, all CATH domains in the same CATH superfamily were grouped together in order to build a BLAST database of the sequences of three-dimensional structures specific to each of the CATH superfamilies.

#### *Generation of the groups of Gene3D sequences represented by the same CATH domain*

In order to assign the closest CATH domain to each of the Gene3D sequences belonging to the same CATH superfamily, we queried each of the sequences in Gene3D against a database of CATH domains in that superfamily using BLAST. The best BLAST hit for each of the Gene3D sequences – provided the identity between the hit and the query was greater than 20% – was considered the closest CATH domain and hence the CATH domain was assigned as the structural representative of this sequence. After performing this classification for the whole set of sequences, all Gene3D sequences represented by the same CATH domain were grouped together and all the sequences within a group considered similar. A total of 2091 unique groups were generated.

#### *Alignment of the CATH domain groups using MUSCLE*

During the previous step of the pipeline, the sequences contained in each of the groups of Gene3D sequences represented by the same CATH domain were considered similar. However, in order to collapse all the mutations from the Gene3D sequences onto the representative CATH domain sequence, the equivalence between pairs of residues needed to be established. To perform this task, multiple sequence alignments were constructed using the alignment package MUSCLE.

#### *Mapping SAAPdb mutations to CATH domain representative structures*

The alignments generated by MUSCLE during the previous step of the pipeline were used to transfer the mutations, both pathogenic (PDs) and neutral (SNPs), from the sequences in Gene3D to the corresponding CATH structural representatives.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

Conceived the idea: AV, JMGI, CO, ACRM. Gathered the data and generated the mapping: JMGI, AB, LM, CY. All authors designed the server and its functionalities. Implemented the server: JMGI. Implemented the database: JMGI. Implemented the webservices: JMGI, AC. Wrote the paper: all authors. All authors read and approved the manuscript. CNIO covered the publication expenses.

## Acknowledgements

The CNIO group is supported by funding from the Consolider BSC (CSD2007-00050) project and the National Institute of Bioinformatics (INB), a platform of 'Genoma España'. Regarding the UCL group, LEMM is funded by a UK Medical Research Council Capacity Building Studentship in Bioinformatics and AB by the Overseas Research Student Awards Scheme and UCL Graduate Research Scholarship. The overall work is part of the common effort under the EMBRACE Network (LSHG-CT-2004-512092). The authors want to thank Jose Manuel Rodriguez and Antonio Rausell, for their help, interesting discussion and ideas.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 8, 2009: Proceedings of the European Conference on Computational Biology (ECCB) 2008 Workshop: Annotation, interpretation and management of mutations. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/10?issue=S8.

## References

1. Collins FS, Brooks LD, Chakravarti A: **A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation.** *Genome Research* 1998, **8(12):**1229-1231.
2. Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY: **Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms.** *Genome Research* 1998, **8(7):**748-754.
3. Schattner P, Diekhans M: **Regions of extreme synonymous codon selection in mammalian genes.** *Nucleic Acids Research* 2006, **34(6):**1700-1710.
4. Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM: **Silent polymorphisms speak: how they affect pharmacogenomics and the treatment of cancer.** *Cancer Research* 2007, **67(20):**9609-9612.
5. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Research* 2001, **29:**308-311.
6. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35(Database issue):**D610-D617.
7. Consortium H: **The International HapMap Project.** *Nature* 2003, **426:**789-796.
8. Bromberg Y, Yachdav G, Rost B: **SNAP predicts effect of mutations on protein function.** *Bioinformatics* 2008, **24(20):**2397-2398.
9. Mort M, Ivanov D, Cooper DN, Chuzhanova NA: **A meta-analysis of nonsense mutations causing human genetic disease.** *Human Mutation* 2008, **29(8):**1037-1047.
10. Torkamani A, Schork NJ: **Accurate prediction of deleterious protein kinase polymorphisms.** *Bioinformatics* 2007, **23(21):**2918-2925.
11. Yue P, Moult J: **Identification and analysis of deleterious human SNPs.** *Journal of Molecular Biology* 2006, **356(5):**1263-1274.
12. Gabdoulline RR, Ulbrich S, Richter S, Wade RC: **ProSAT2-Protein Structure Annotation Server.** *Nucleic Acids Res* 2006, **34(Web Server):**W79-83.

13. Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M: **PMUT: a web-based tool for the annotation of pathological mutations on proteins.** *Bioinformatics* 2005, **21(14):**3176-3178.
14. Ferrer-Costa C, Orozco M, de la Cruz X: **Sequence-based prediction of pathological mutations.** *Proteins* 2004, **57(4):**811-819.
15. Wang Z, Moult J: **Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain.** *Proteins* 2003, **53(3):**748-757.
16. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Research* 2003, **31(13):**3812-3814.
17. Ferrer-Costa C, Orozco M, de La: **Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties.** *Journal of Molecular Biology* 2002, **315(4):**771-786.
18. Wang Z, Moult J: **SNPs, protein structure, and disease.** *Human Mutation* 2001, **17(4):**263-270.
19. Hurst JM, McMillan LEM, Porter CT, Allen J, Fakorede A, Martin ACR: **SAAPdb web resource: a large scale structural analysis of mutant proteins.** *Human Mutation* 2009 in press.
20. Sayers EWW, Barrett T, Benson DAA, Bryant SHH, Canese K, Chetvernin V, Church DMM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LYY, Helmberg W, Kapustin Y, Landsman D, Lipman DJJ, Madden TLL, Maglott DRR, Miller V, Mizrachi I, Ostell J, Pruitt KDD, Schuler GDD, Sequeira E, Sherry STT, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TAA, Wagner L, Yaschenko E, Ye J: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2009, **37(Database issue):**D5-D15.
21. Claustres M, Horaitis O, Vanevski M, Cotton RG: **Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases.** *Genome Research* 2002, **12(5):**680-688.
22. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA: **The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution.** *Nucleic Acids Research* 2007:D291-D297.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Research* 2000, **28:**235-242.
24. Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, Orengo C: **Gene3D: comprehensive structural and functional annotation of genomes.** *Nucleic Acids Research* 2008:D414-D418.
25. Consortium U: **The Universal Protein Resource (UniProt).** *Nucleic Acids Research* 2007:D193-D197.
26. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997, **25(17):**3389-3402.
27. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5:**113.
28. Laskowski RA, Chistyakov VV, Thornton JM: **PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids.** *Nucleic Acids Research* 2005:D266-D268.
29. Loeys BL, Chen J, Neptune ER, Judge DP, Podowski M, Holm T, Meyers J, Leitch CC, Katsanis N, Sharifi N, Xu LL, Myers LA, Spevak PJ, Cameron DE, De Backer JD, Hellemans J, Chen Y, Davis EC, Webb CL, Kress W, Coucke P, Rifkin DB, De Paepe AMD, Dietz HC: **A syndrome of altered cardiovascular, craniofacial, neurocognitive and skeletal development caused by mutations in TGFBR1 or TGFBR2.** *Nature Genetics* 2005, **37(3):**275-281.
30. Uzun A, Leslin CM, Abyzov A, Ilyin V: **Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways.** *Nucleic Acids Research* 2007, **35:**W384-W392.
31. Izarzugaza JMG, Redfern OC, Orengo CA, Valencia A: **Cancer associated mutations are preferentially distributed in protein kinase functional sites.** 2009 in press.
32. Izarzugaza JMG, Redfern OC, Orengo CA, Valencia A: **Distribution of pathogenic mutations within the representative structures in the CATH hierarchy.** 2009 in press.
33. Piirilä H, Väliaho J, Vihinen M: **Immunodeficiency mutation databases (IDbases).** *Human Mutation* 2006, **27(12):**1200-1208.
34. Kwok CJ, Martin ACR, Au SWN, Lam VMS: **G6PDdb, an Integrated Database of Glucose-6-phosphate Dehydrogenase (G6PD) Mutations.** *Hum Mutat* 2002, **19:**217-224.
35. Kemball-Cook G, Tuddenham E, Wacey A: **The factor VIII Structure and Mutation Resource Site: HAMSTeRS version 4.** *Nucl Acids Res* 1998, **26:**216-219.
36. Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M: **Impact of Mutant p53 Functional Properties on TP53 Mutation Patterns and Tumor Phenotype: Lessons from Recent Developments in the IARC TP53 Database.** *Hum Mutat* 2007, **28:**622-629.
37. Leigh SEA, Foster AH, Whittall RA, Hubbart CS, Humphries SE: **Update and Analysis of the University College London low Density Lipoprotein Receptor Familial Hypercholesterolemia Database.** *Ann Hum Genet* 2008, **72:**485-498.
38. Tuchman M, Jaleel N, Morizono H, Sheehy L, Lynch MG: **Mutations and Polymorphisms in the Human Ornithine Transcarbamylase gene.** *Hum Mutat* 2002, **19:**93-107.
39. Wroe R, Wai-Ling Butler A, Andersen PM, Powell JF, Al-Chalabi A: **ALSOD: the Amyotrophic Lateral Sclerosis Online Database.** *Amyotroph Lateral Scler* 2008, **9:**249-250.
40. Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, Orengo C: **Gene3D: Comprehensive Structural and Functional Annotation of Genomes.** *Nucleic Acids Res* 2008, **36:**D414-D418.
41. Rattei T, Tischler P, Arnold R, Hamberger F, Krebs J, Krumsiek J, Wachinger B, Stümpen V, Mewes W: **SIMAP-structuring the Network of Protein Similarities.** *Nucleic Acids Res* 2008, **36:**D289-D292.

BMC Bioinformatics

BioMed Central

Research

Open Access

# Extraction of human kinase mutations from literature, databases and genotyping studies

Martin Krallinger*†[1], Jose MG Izarzugaza†[1], Carlos Rodriguez-Penagos[2] and Alfonso Valencia[1]

Address: [1]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre, Madrid, Spain and [2]Barcelona Media, Centre d'Innovació, Av. Diagonal 177, Barcelona, Spain

Email: Martin Krallinger* - mkrallinger@cnio.es; Jose MG Izarzugaza - jmgonzalez@cnio.es; Carlos Rodriguez-Penagos - carlos.rodriguez@barcelonamedia.org; Alfonso Valencia - valencia@cnio.es

* Corresponding author †Equal contributors

*from* ECCB 2008 Workshop: Annotations, interpretation and management of mutations (AIMM)
Cagliari, Italy. 22 September 2008

Published: 27 August 2009

*BMC Bioinformatics* 2009, **10**(Suppl 8):S1 doi:10.1186/1471-2105-10-S8-S1

This article is available from: http://www.biomedcentral.com/1471-2105/10/S8/S1



## Abstract

**Background:** There is a considerable interest in characterizing the biological role of specific protein residue substitutions through mutagenesis experiments. Additionally, recent efforts related to the detection of disease-associated SNPs motivated both the manual annotation, as well as the automatic extraction, of naturally occurring sequence variations from the literature, especially for protein families that play a significant role in signaling processes such as kinases. Systematic integration and comparison of kinase mutation information from multiple sources, covering literature, manual annotation databases and large-scale experiments can result in a more comprehensive view of functional, structural and disease associated aspects of protein sequence variants. Previously published mutation extraction approaches did not sufficiently distinguish between two fundamentally different variation origin categories, namely natural occurring and induced mutations generated through in vitro experiments.

**Results:** We present a literature mining pipeline for the automatic extraction and disambiguation of single-point mutation mentions from both abstracts as well as full text articles, followed by a sequence validation check to link mutations to their corresponding kinase protein sequences. Each mutation is scored according to whether it corresponds to an induced mutation or a natural sequence variant. We were able to provide direct literature links for a considerable fraction of previously annotated kinase mutations, enabling thus more efficient interpretation of their biological characterization and experimental context. In order to test the capabilities of the presented pipeline, the mutations in the protein kinase domain of the kinase family were analyzed. Using our literature extraction system, we were able to recover a total of 643 mutations-protein associations from PubMed abstracts and 6,970 from a large collection of full text articles. When compared to state-of-the-art annotation databases and high throughput genotyping studies, the mutation mentions extracted from the literature overlap to a good extent with the existing

knowledgebases, whereas the remaining mentions suggest new mutation records that were not previously annotated in the databases.

**Conclusion:** Using the proposed residue disambiguation and classification approach, we were able to differentiate between natural variant and mutagenesis types of mutations with an accuracy of 93.88. The resulting system is useful for constructing a Gold Standard set of mutations extracted from the literature by human experts with minimal manual curation effort, providing direct pointers to relevant evidence sentences. Our system is able to recover mutations from the literature that are not present in state-of-the-art databases. Human expert manual validation of a subset of the literature extracted mutations conducted on 100 mutations from PubMed abstracts highlights that almost three quarters (72%) of the extracted mutations turned out to be correct, and more than half of these had not been previously annotated in databases.

## Background

Protein kinases are the most ubiquitous family of signaling molecules in human cells, accounting for approximately 2% of the proteins encoded by the human genome [1]. They can be further divided into sub-families that share significant similarity both at the sequence and structural level. A common feature of all kinases is their ability to transfer the terminal phosphate of ATP to serine, threonine or tyrosine residues of a target protein. Empirical studies also suggest a common catalytic mechanism whereby ATP and active site divalent cations are bound as well, and phospho-transfer is carried out by a shared set of amino acids. Despite these functional commonalities, experiments in yeast models [1,2] suggest that the protein kinase family as a whole is highly promiscuous, phosphorylating a range of different protein substrates, although individual sub-families may display a remarkable substrate specificity [3]. Kinases have a domain committed to the general function of catalysis, while another region (or even regions) are used in many cases to confer substrate specificity to the enzyme, without altering the general kinase folding, interfering with ATP binding or the general reaction mechanism. For reviews on the evolution of kinase structure and function see [4-7]. Several efforts have been made to provide a comprehensive access to information relevant to characterize human protein kinases through specific databases such as KinBase [1], KinMutBase [8] and MoKCa [9], or more general databases like PDB [10], PFAM [11] and CATH [12], storing information important to understand disease-association, functional and structural properties of kinases.

The relation of kinases with a number of diseases [13] and in particular with cancer [14-16] has prompted a number of large scale studies, in particular, Greenman *et al.* carried out the first large scale study of the variation associated with 518 protein kinase genes in 210 samples of cancer tissues and cell lines. Other HT studies not specifically restricted to kinases have obviously also contributed in understanding and providing information on mutations in protein kinases [14,15]. The interest of kinases and their implication in disease processes has continued with the study of Sjöblom and colleagues [15]. For a detailed review refer to Baudot *et al.* [17].

The sizeable amounts of information provided by large scale variation studies and the growing efforts of databases and resources to store and curate this information, are still not perfectly/completely connected with the many efforts dedicated to the detailed study of specific kinases in various biological systems [18] published in individual research papers. For instance, it is still a challenging task to establish for which individual mutations detected in HT studies there is already available information in the literature. Manual inspection and curation of specific variation studies and the exact linking to the textual information requires considerable resources.

Despite difficulties in extracting more complex language expressions referring to mutation mentions, regularities in describing mutations based on existing nomenclature conventions, promoted the implementation of automated information extraction and text mining systems for the identification of mutations in the literature [19-30]. Table 1 provides a short summary of previously published literature mining efforts for mutation information extraction.

Even though most of the existing manually curated mutation annotation resources are based on reading full text articles, existing automated systems mainly relied only on (subsets of) PubMed abstracts or a small collection of full text articles. To facilitate the interpretation of the biological implications and phenotypic effects of a given mutation, not only by clinical experts but also by database curators or for designing biochemical experiments (drug design and molecular functional studies) it is crucial to know whether a given mutation has been experimentally generated or is present in a naturally occurring sequence variation. This aspect has generally been neglected by previously developed approaches. Finally, only few systems were able to show results based on the combination of

heterogeneous data derived from multiple information sources, derived from literature as well as based on experimental data generated by genotyping studies.

Here we examined the use of text mining methods to extract information from the literature about protein kinases and their specific mutations, link this information to the corresponding protein sequences from databases (normalization) and analyze how this information is distributed in protein kinases related databases and repositories. The results of comparing information from databases and text repositories are analyzed in terms of the quality of the information provided and the significance in terms of the knowledge related to PK structure and function. We therefore applied an available mutation extraction system, called MutationFinder [23] to the whole collection of abstracts contained in PubMed database as well as to a large set of automatically retrieved full text articles. To determine if a putative mutation mention really corresponds to a mutation or actually to something else we developed a module that allows filtering of false positive mutation mentions through a combination of named entity recognition, dictionary look-up and rule based methods. A supervised machine learning method relying on the SVM algorithm was used to score and classify based on its context whether a given mutation mention correspond to an experimentally generated (induced) mutation or is a natural sequence variant. A protein mention normalization system together with an mutation sequence checking approach was used to detect associations between mutations and human kinases co-cited in the literature. Validation and comparison to multiple existing mutation resources, including the SwissProt database [31] as well as the COSMIC [32], Greenman/Wood dataset of somatic mutations [14,16], KinMutBase [8], and SAAPdb databases [33]. The extraction and comparative mutation analysis were followed by a structural examination of the distribution of the different type of mutations in kinase regions of structural and functional importance.

## Results and discussion

Here we present a workflow for extracting mutations within human protein kinases. The pipeline integrates article retrieval, detection of mutations mentioned in the literature, and a final validation of mutations linked to their corresponding protein sources. We carried out a comparative analysis of multiple annotation resources containing different mutation types. An overview of the resulting approach is presented in figure 1, illustrating the main steps of the mutation extraction pipeline.

### *Systematic extraction and disambiguation of mutation mentions*

For the initial extraction of single amino acid substitutions we applied the MutationFinder system, a modular software for point mutation recognition based on regular expressions and patterns detecting mutation mentions corresponding to residue abbreviations as well as other language expressions used to describe mutation events [34]. This system shows a competitive performance in terms of recall/precision when compared to other strategies [28] and has been evaluated using a manually generated Gold Standard collection of abstracts [23].

We applied the MutationFinder tool to the whole PubMed database (November 2008), resulting in the detection of 302,956 mutation mentions from 88,405 records, corresponding to a total of 61,329 unique mutation types (i.e. wild type residue, sequence position and mutant residue triplets). A more detailed analysis of the most frequent mutation types (see additional file 2), illustrated the importance of the Cystein to Tyrosine mutation in position 282 (C282Y, corresponding to the dbSNP:rs1800562) and the Histidine to Aspartic Acid mutation at position 63 (H63D, dbSNP:rs1799945), both occurring in the hereditary hemochromatosis protein (HFE, SwissProt:Q30201), known to be associated to several human diseases. These two mutations are mentioned over 3,500 and 1,900 times respectively. Some of the most frequent mutation types corresponded to cases of false positive (ambiguous) mutations mentions that actually consisted in names of cell lines (T47D cells) or mouse strains (G93A mouse model).

As the MutationFinder system was developed and evaluated using a collection of abstracts known to be relevant for mutations, primarily derived from citations related to mutant protein structures from the Protein Data Bank (PDB), we carried out a coarse level consistency analysis to determine how scalable this system is when applied to the whole PubMed database, where many articles do not necessarily resemble the data collection used for the initial system development. Assuming that the overall mutation types, contained in manually annotated resources like the SwissProt database should be similar to the ones encountered throughout PubMed we compared mutations extracted automatically from the literature to information contained in SwissProt, namely mutations being annotated as either natural variant, induced (mutagenesis) or both single amino acid substitutions. A comparative analysis of the frequencies of annotated mutation pairs (wild type residue and the associated mutation) showed that there are considerable differences between the mutations often encountered in naturally occurring variations as opposed to experimentally induced amino acid changes. The overall profiles resulting from the relative percentages

**Table 1: Literature mining approaches for mutation extraction. The additional materials sections (Additional file 1) provides a more detailed description of each method.**

| Collection of existing mutation extraction approaches | | |
|---|---|---|
| **Method** | **Main characteristics and descriptive keywords of the approach** | **Ref** |
| MEMA | Uses regular expressions, gene and protein mention detection, co-mention proximity, OMIM validation | [19] |
| MuteXt | Uses regular expressions, GPCR and NR mention detection, co-mention proximity, sequence check | [27] |
| Yip *et al.* | Uses regular expressions, protein mention detection, SwissProt validation, extensive sequence check | [28] |
| CoagMDB | Uses regular expressions, serine protease mention detection, sequence check | [41] |
| Mutation GraB | Uses regular expressions, protein mention detection, graph shorted distance, sequence check | [20] |
| Mutation Miner | Uses regular expressions, protein mention detection, sentence co-mention | [21] |
| MuGeX | Uses regular expressions, protein mention detection, protein and DNA mutation disambiguation | [24] |
| VTag | Machine learning (CRF) detection of acquired sequence variations mentions (mutations, translocations, deletions) | [26] |
| OSIRISv1.2 | Detection of human gene variations corresponding to SNPs | [42] |
| MutationFinder | Uses regular expressions and patterns, protein mutations, complex language expressions | [23] |

of mutation types from different sources are similar (see figure 2A), despite existing differences encountered in terms of the most frequent mutations pairs extracted through text mining when compared to manual annotations (figure 2B).

The most remarkable differences in case of the relative frequencies of mutation types extracted from PubMed compared to SwissProt corresponded to mutations formed by the residues, A, T, C and G (i.e. G-T, C-T, A-G and T-A substitutions). This is due to the intrinsic ambiguity of single letter mutation mentions that can correspond to both mutations at the DNA or protein level depending on the context. In order to distinguish between these two mutation levels, additional processing would be required. A more detailed analysis of the wild type residue and the mutant residue frequencies revealed that automatically extracted mutation residues are in line from what would be expected when examining the relative frequencies within a manually curated database (figure 3A and 3B).

The most frequently mutated residues mentioned in PubMed are Arginine, Glycine and Serine, corresponding also to the top ranking ones annotated in SwissProt. SwissProt shows more variability when comparing annotations of wild type residues from naturally to induced variations. In case of experimentally generated mutations, the residues most frequently replaced are Serines, Lysines, Arginines, Cysteines and Tyrosines, corresponding to functionally important residues. Considering the mutant residues, the literature mining extracted residues are consistent with the mutant residues from SwissProt, which shows great variation in case of Alanine induced and natural variant mutants. This can be explained by the widespread use of experimental approaches relying on the Alanine scanning method for identifying functionally relevant sites, as substitutions to Alanine usually still allow protein folding yet may give an altered phenotype. A more detailed description of the mutation disambiguation and filtering approach to remove false positive mutation mentions is provided in the method section.

### *Scoring variation origin categories: artificial and natural mutations*

For extraction and management of biological annotations and to carry out functional analysis of mutations it is crucial to know the level of granularity and experimental context used for determining the phenotypic effect of a given amino acid substitution. The SwissProt database distinguishes here between induced or artificial (mutagenesis)

and natural variant mutations, corresponding the former to less than 13 percent of the mutation annotations. Characterizations at the level of molecular functional implications and sub-cellular interactions of specific residues are commonly studied through experimentally induced amino acid changes. On the other hand associations to diseases such as certain cancer types and relevance for population groups or patients of a given mutation is usually studied by examining naturally occurring sequence variants. To address this important issue, allowing mutation mention scoring for each of these two basic categories of phenotypic descriptions we applied a supervised machine learning strategy for mutation sentence classification. We applied a SVM algorithm (using radial basis function as kernel) trained on a balanced sample set of 3,482 (71%) labeled sentences for induced and natural variant mutations and evaluated it on an independent test set of 1,400 sentences (29%), obtaining an accuracy of 94.64 (recall: 94.57 and precision of 94.71) on this collection. The size of the feature dictionary used by the classifier was of 11,803 word types (unique words, not stemmed). A manual inspection of the generated feature dictionary revealed that some of the relevant features corresponded to terms comprised in experimental techniques used to generate artificial mutations, such as site-directed mutagenesis. This basic evaluation schema is suitable to determine the performance on a controlled set of balanced instances, but does not take into account the actual distribution of the classes within a large collection of unlabelled data nor cases that even by human experts can not be clearly classified into one of these binary categories. Therefore we carried out both, a classification consistency analysis on the resulting sentence scores as well as a detailed evaluation and comparison against manually examined mutation sentences (see figures 4A to 4F).

To determine the overall classification and scoring consistency on the level of the whole mutation sentence collection extracted from PubMed described in the previous subsections, we analyzed the distribution of a database confirmed set of natural variant mutations against a randomly chosen set of mutation mentions. The first collection of sentences corresponded to mutation mentions cross-checked from SwissProt annotations as natural variants, resulting in a total of 10,886 sentences, none of these were contained in the original training nor test set. The second collection was constructed by randomly selecting an equal number of mutation mentioning sentences from PubMed abstracts. For each of these two sets we determined the corresponding sentence scores generated by the classifier. Figure 4A shows the box plot of the sentence classifier scores for the 10,886 natural variant mutation sentences and the equally sized random collection of mutation mentioning sentences. The scores of natural variant mutation sentences (mean of 1.37) were significantly higher when compared to the random subset (mean of 0.09), indicating that the overall scoring of natural variant mutation mentioning sentences derived from the independent SwissProt annotations are consistently higher than a random subset.

For practical purposes it is often useful to determine the actual performance of a system for a discrete set of score intervals or cut-offs, to enable a more efficient selection of instances for further examination or manual curation. Therefore we selected random subsets for sentence score intervals ranking from above 4 (positive class, natural variant relevant) to minus 4 (negative class, induced mutation or mutagenesis). Sentences of each of these sets underwent a two-step blindfold manual classification process to provide a more fine-grained analysis of the different aspects that might influence the actual systems performance. The first step consisted in classifying whether the sentence is mentioning a mutation or not to determine the effect of false positive mutation extraction. As a separate class we also recorded cases of mutation mentioning sentences where the directionality of the automatically extracted mutation event (wild type residue vs. mutant residue) was wrongly derived. When considering the mutation extraction performance across the score intervals for mutations classified as induced and natural variants, it seems that it was more difficult to correctly identify mutation mentions in abstracts that where close to the classification boundary or where scored as experimentally generated mutations.

The second step involved manually classifying mutation mentions into one of the following categories: (1) natural variant, (2) induced mutation or (3) unclear cases. We decided to add the latter category to take into account mutation mentions that even by humans could not be classified clearly into one of the two other types, either because the context of mention is not informative enough or because it is a truly ambiguous case. Figure 4E and 4F provide a detailed overview of the results obtained from this multi-step manual mutation classification for each of the score intervals. The classifier results are identical to human classification in over 95% of the cases for very high and very low mutation sentences score intervals, but drop to less than 77% for cases where mutations where classified as natural variants with classifier scores close to the classifier decision boundary (score interval of 1 to 0). Finally we also selected randomly a larger set of mutation mentions and carried out a bind fold manual labeling of these cases, to analyze the overall performance of the mutation classifier as well as to determine the distribution of natural variant and induced mutation mentions from PubMed abstracts. From the initially extracted set of mutations, 93.9% corresponded to correct mutation mentions without applying the mutation filter, compared to 97.0%

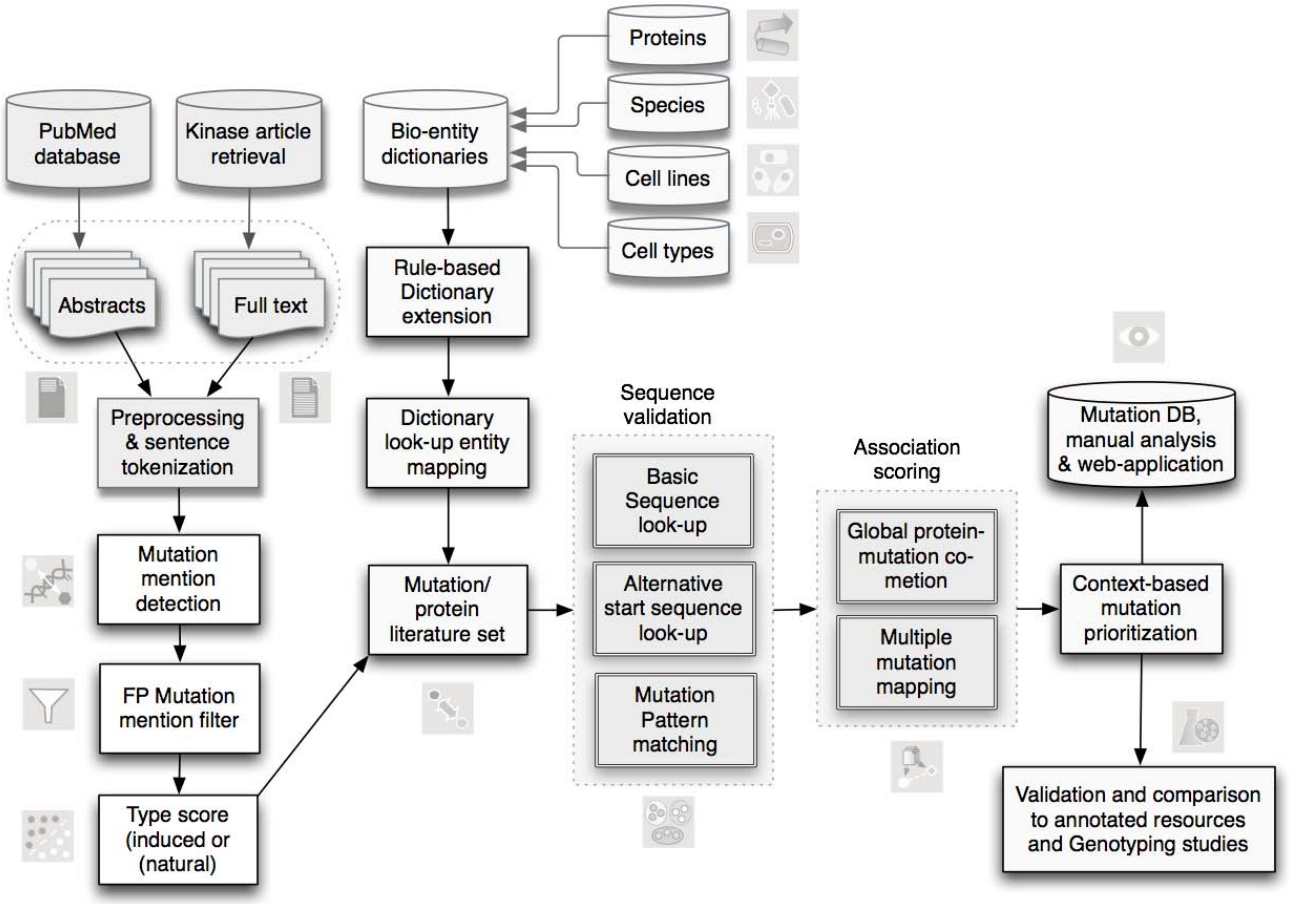



**Figure 1**

**Flow chart of the presented literature mining approach for mutation extraction**. This flow chart provides an overview of the different processing steps to extract mutations relevant for human kinases. The main steps include the construction of a kinase relevant article collection, the detection of mutation mentions, the scoring of the type of mutation (induced/natural variant), the linking of mutations to the corresponding protein sequence and the comparison to existing databases.

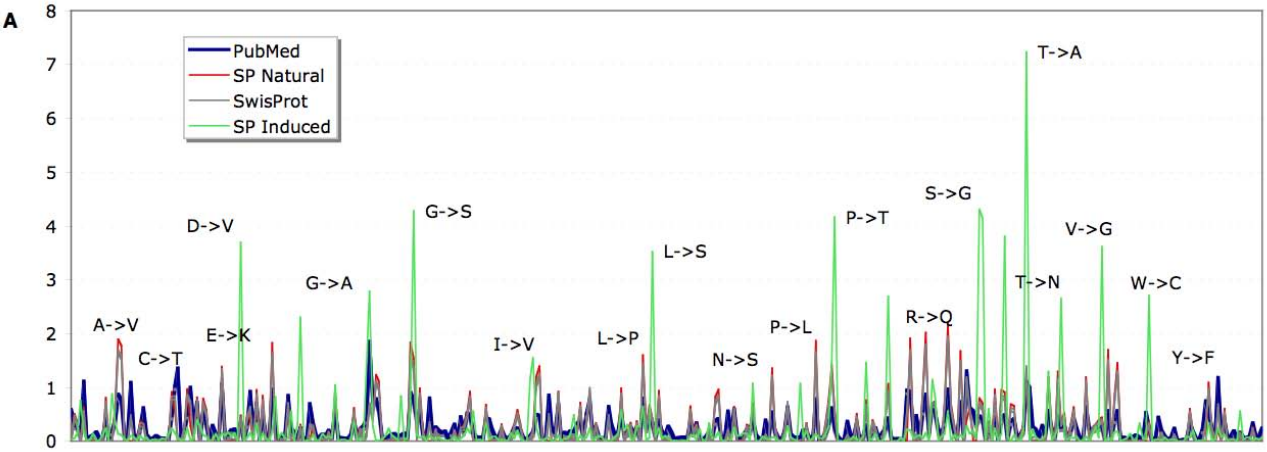
when applying the false positive mutation detection step. Out of the correctly extracted mutations, 49.75% were manually classified as natural variant mutations, very close to the 47.2% of mutations classified as induced mutations. Surprisingly only 3.05% corresponded to unclear mutation types. Evaluating the sentence classifier results against the manually classified labels resulted in a precision of 93.88%, a recall of 91.09% (balanced f-score of 92.46) and an accuracy of 93.88, in line with the performance obtained with the previously used test set. An interesting false negative case was the sentence: The hexameric structure is important for protein stability, as demonstrated by studies with natural mutants (the Killer-of-prune mutant of Drosophila NDP kinase and the S120G mutant of the human NDP kinase A in neuroblastomas) and with mutants obtained by site-directed mutagenesis. In this particular case the authors refer to both natural variants as well as induced mutations, being S120G actually a natural mutation.

#### *Linking mutation mentions to human kinase sequences*

Providing associations of mutations to their corresponding protein record and sequence is crucial to facilitate a more detailed characterization of structural effects of a given mutation and distribution within certain protein domains. This also allows direct comparison to functional annotations of proteins and mutations contained in manually curated annotation databases as well as to large-scale experimental results obtained by genotyping studies. Here we focus on associating the extracted mutations specifically to human protein kinases.

To obtain links between mutation mentions and human kinases we assumed that the corresponding protein

A




B

| Mutation | PubMed | Mutation | SP Natural | Mutation | SP Induced |
|---|---|---|---|---|---|
| *G->A* | 1.85 | *R->Q* | 2.16 | *T->A* | 7.23 |
| *C->T* | 1.35 | *R->H* | 2.02 | *S->F* | 4.31 |
| *S->A* | 1.30 | *R->C* | 1.91 | *G->S* | 4.28 |
| *Y->F* | 1.18 | *A->T* | 1.90 | *P->T* | 4.16 |
| *A->G* | 1.11 | *P->L* | 1.87 | *S->G* | 4.14 |
| *T->A* | 1.10 | *G->R* | 1.83 | *S->P* | 3.81 |
| *C->A* | 1.09 | *E->K* | 1.83 | *D->V* | 3.69 |
| *D->N* | 1.06 | *A->V* | 1.77 | *V->G* | 3.61 |
| *C->S* | 1.02 | *V->I* | 1.70 | *L->S* | 3.52 |

C

| Source | Values of Correlation (rs) |
|---|---|
| SwisProt | 0.908673111 |
| SP Natural | 0.787173605 |
| SP Induced | 0.591905892 |

**Figure 2**
**Mutation type frequencies from PubMed and SwissProt**. A. Relative frequency of each mutation type derived from PubMed abstracts and from the SwissProt database. B. Most frequent mutation types from PubMed abstracts, and from SwissProt (SP), annotated as natural variant or induced (mutagenesis) substitutions. C. Values of the Spearman rank correlation between the text mining derived mutation types and the database derived mutation types. All p-values are below 10e-6, therefore statistically significant.

names should be co-mentioned in the articles. After extracting all the mutation mentions from PubMed abstracts and a large collection of full text articles, these two data sets were processed for retrieving mentions of human protein kinases. In order to detect kinase protein mentions we applied a dictionary look-up approach, similar to strategies that participated successfully at the gene normalization task of BioCreative II [35]. To take into account inter- and intra-species protein name ambiguity, rather than using very strict protein-organism source co-mention criteria based on relative textual distances, we calculated for each article two scores reflecting (1) the contextual similarity of the article to the SwissProt protein record and (2) the overall association of the article to human species terms from the total set of tagged species terms.

This high recall protein normalization scoring strategy was followed by a more stringent sequence validation approach that allowed us to detect links of mutations and proteins by checking whether the actual mutation mention can be confirmed by looking them up at the protein sequence position. We restricted our analysis specifically to mutations occurring in the protein kinase domain, as defined in Kinbase [1]. A total of 567 triplets (i.e. article-mutation-protein associations) derived from abstracts could be validated by checking whether the extracted wild type residue was found at the mutated position in the protein sequence. In addition to this basic sequence look-up validation method we implemented five complementary mutation-sequence mapping strategies that take into account both, errors resulting from the wrong detection of the actual directionality of the extracted mutation with respect to wild type and mutant residues as well as inconsistencies and alternative sequence counting between the article and the database kinase sequence (see methods section). By applying this additional matching strategies we were able to recover 437 additional hits, corresponding to 43.53 percent of the total set of sequences validated protein mutation pairs. This resulted in a total of 1,004 triplets from 714 abstracts. In case of full text articles, the total number of triplets detected by the basic mapping was 3,911, being another 3,917 triplets recovered through additional sequence mapping methods. This resulted in 7,828 triplets from 3,496 full text articles. The average number of sequence validated mutations in the Protein

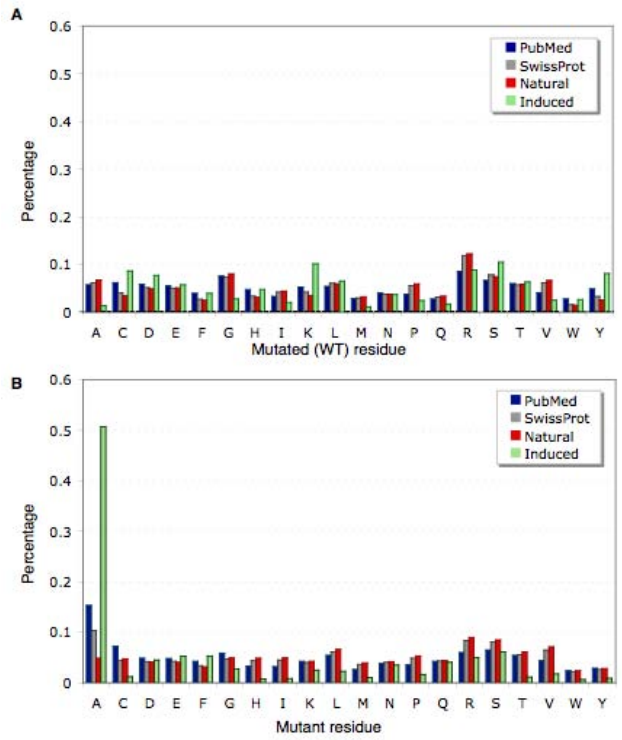



**Figure 3**
**Comparative analysis of wild type residues and mutations extracted from SwissProt and using text mining**. This chart illustrates differences in terms of wild type and mutant residue frequencies derived from the SwissProt database and obtained through automatic literature processing. A. Relative frequency of each wild type residue derived from mutations extracted from PubMed abstracts and from the SwissProt database. B. Relative frequency of each mutant residue derived from mutations extracted from PubMed abstracts and from the SwissProt database.

Kinase Domain for abstracts was 1.41 and for full text articles 2.24, implying that often more than a single mutation is described in a given paper.

The global context of co-mention of kinase proteins and mutations defined by the multi-document collection where these co-occur, can be indicative for the actual importance of a particular mutation, being described and studied in various different paper. To use information provided by the corpus co-mention context, in addition to the total number of documents where the sequence validated mutation-kinase pair co-occurred we calculated the mutual information for each mutation pair.

### *Comparison of text mining mutations to databases and genotyping studies*

Several genomic studies (including a comprehensive analysis of all human kinases) have been dedicated to the characterization of mutations occurring in protein kinases in a variety of cancer tissues and cell lines. In these studies, a number of point mutations detected in somatic cell lines have been found to be associated with specific cancer types. The pathogenicity of these mutations depends on multiple factors related to the complex molecular environment in which protein kinase function takes place. Of special interest are mutations found within the protein kinase domain, as it is essential for the functional activity of these proteins. We therefore focused our analysis on automatic extraction of mutations mapped to this particular domain, common to all kinases, and carefully examined how they relate to previously characterized mutations retrieved from multiple databases and experimental high throughput genomic studies. We used the kinase domain definition followed by Kinbase [1], analyzing both the distribution of mutations within the Protein Kinase Domain, as well as the distribution of mutations according to the corresponding protein family topology. A total of 643 kinase domain sequence mutations were extracted from PubMed abstracts for a total of 128 different proteins. When considering the full text collection, we were able to increase considerable this number, obtaining a total of 6,970 mutation-protein pairs from 325 proteins. Using full text articles resulted therefore in a considerable increase of recovered mutations (more than 10 times more mutation-protein pairs when compared to abstracts) as well as being useful to increase the recall of proteins for which mutations had been extracted (more than doubling the initial number derived form abstracts alone). The increased recall for full text papers clearly justifies the computational effort required to retrieve and preprocess them.

Figure 5A shows the distribution of the mutations extracted from the literature into the different groups in which Kinbase [1] classifies the protein kinase domain of the human kinases. For a more detailed description of the different kinase groups refer to the methods section. Although there are differences in the number of mutations present – with more than a half of the mutations either within the TK or the CMGC clades – all the main groups are represented in the results both when PubMed abstracts or full text articles are taken into consideration. Figure 5B depicts the normalized distribution of mutations in the different protein kinase domain groups in which Kinbase classifies the human kinome when the abstracts and the full text articles are taken into account respectively. It is evident that no matter which dataset is used, either abstracts or full-text articles, the distributions are very similar independently of the very uneven absolute numbers between both datasets, which confirms that complementary results are provided by the two very different approaches.

*Confirmation and comparison to experimental and curated data*

In order to assess whether the mutations recovered from the existing literature by our system were already present in commonly used databases or are newly recovered instances, we herein studied the overlap between the mutations in the protein kinase domain both in the databases and the results from our extraction pipeline. Table 2 represents the percentage of each database covered by the Text Mining results.

*Recovery of disease associated mutations: overlap with KinMutBase*

KinMutBase [8] is a manually curated knowledge base for human disease-related mutations in protein kinase domains. At the time of the study, March 2009, a total of 83 single-point pathogenic mutations in the protein kinase domain of 10 different proteins were provided in KinMutBase. We were able to confirm 32 (38.55%) of these mutations from abstracts and also the same number from full text articles. When combining the mutations from both article collections we recover more than half of the mutations from KinMutBase, namely a total of 43 mutations (51.81%). This suggests that the mutation mentions from both document collections are essentially complementary, and some of them could only be detected in one of the document sets.

*Recovery of natural variant and induced mutations: overlap with SwissProt database*

The Swissprot Variant database [31] provides experimentally-verified information about mutations present in UniProtKB, containing a set of 710 mutations in the protein kinase domain of 194 different proteins. 251 (35.35%) of them corresponding to mutagenesis experiments whereas 459 (64.65%) correspond to reported natural sequence variants. Using our text mining approach we were able to recover 134 (18.87%), 328 (46.20%) and 365 (51.41%) of the mutation contained in the database when the abstracts, the full texts and the combination of both was used, respectively.

When considering the overlap of extracted mutations with respect to each of these two mutation type classes (natural variants and mutagenesis) we were able to obtain similar percentages for both groups from the combined article collection, 50.11% of the mutations annotated as natural variant and 53.78% of the mutations annotated in SwissProt as mutagenesis.

Interestingly, we found differences in the overlap percentages of recovered mutations from abstracts and from full text articles when looking at these mutation classes individually. When considering the mutations derived from abstracts, 21.57% of the natural variant annotated mutations could be detected, as opposed to only 13.94% of the mutagenesis annotated mutations. The opposite trend was observed in case of full text articles, where we extracted 52.59% of the induced mutations and 42.70% of the natural variant mutations. This suggests to certain extent that experimentally induced mutations annotated in SwissProt are usually not mentioned in abstracts, but rather in full text articles.

*Recovery of structurally important mutations: overlap with the SAAPdb repository*

The SAAPdb [33] is a resource for the analysis and visualization of the structural effects of mutations. At the time of this study, SAAPdb contained 610 point mutations located in the protein kinase domain of 230 proteins. 52.95% (323) of the information corresponds to mutations previously reported as pathogenic deviations (PDs) whereas the rest corresponds to neutral SNPs (287, 47.05%).

Our system recovered 65 (10.66%) and 106 (17.38%) of the mutations previously stored in the database when the abstracts and full text articles were taken into consideration. For the joint abstracts-fulltext dataset, 125 (20.49%) of the mutations present in SAAPdb were found.

With regard to the pathogenicity of the mutations found, for the particular case of the combined dataset, we were able to find 123 (38.08%) of the pathogenic deviations, whereas only 2 neutral SNPs were recovered. This highlights the fact that the literature is biased towards those mutations known to be functionally active and harmful for the individual. It is interesting to remark that none of the other databases analyzed contained records for the neutral SNPs in SAAPdb either.

*Recovery of somatic mutations: overlap with the Greenman and Wood dataset*

The Greenman and Wood dataset was built from the results shown in the original papers [14,16] by the authors where they report 254 somatic mutations corresponding to the protein kinase domain of 164 proteins in diverse human cancers. In addition, the mutations are sub-classified according to the pathogenic character predicted into drivers (cancer associated somatic mutations) and passengers (neutral mutations). The contribution of each class to the whole database is 46.85% and 53.15% respectively.

Our system recovers only a very small fraction of these somatic mutations since only 13 (5.12%) of the instances in the dataset were able to be recovered in the best case scenario, where the combined article set (abstracts+fulltext) was used. This means that only a small proportion of the mutation dataset detected by experimental High Throughput approaches could be linked directly to other

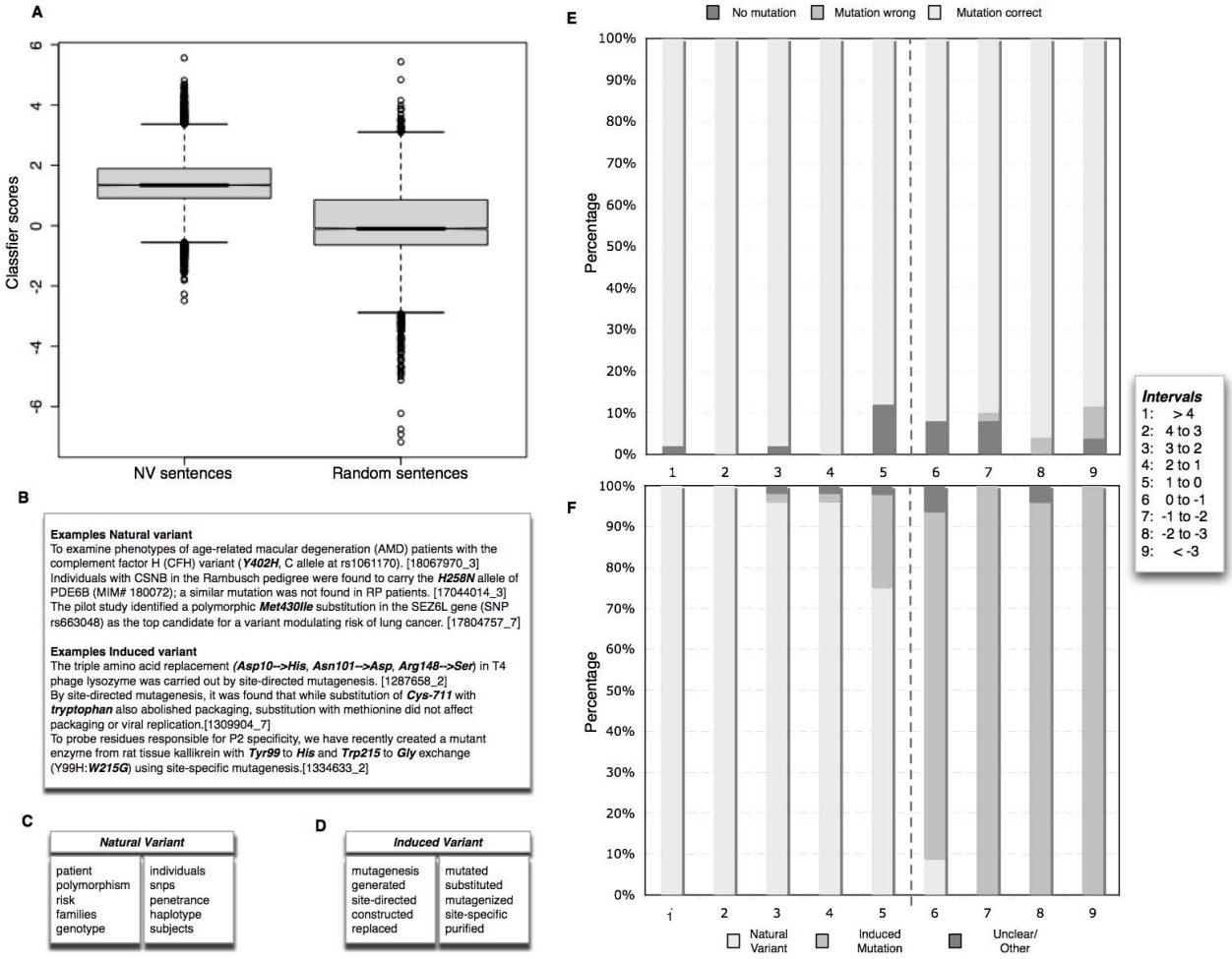



**Figure 4**

**Evaluation of classifying induced mutation mentions and natural variants**. A. Box plot of the sentence classifier scores for Natural Variant (NV) annotated mutations in SwissProt and a random subset of sentence scores from mutation mentioning sentences. B. Example cases of mutation mentions corresponding to natural variant and induced mutations. C. Example features used by the sentence classifier for the positive class (Natural Variant) and the Negative class (Figure D, induced mutation). E and F Manual classification result for 50 randomly selected mutation mentioning sentences for classifier score intervals. (1) Score above 4, (2) score range of 4-3, (3) score range 3-2, (4) score range 2-1, (5) score range 1-0, (6) score range 0 to minus 1, (7) score range from minus 1 to minus 2, (8) score range from minus 2 to minus 3, (9) score range below minus 3. Positive scores correspond to mutations classified as natural variant, negative scores correspond to mutations classified as induced/mutagenesis.

literature evidences. Our system recovered 9 (7.56%) driver mutations versus 4 (2.96%) passenger mutations. A very similar trend was observed for the case of the COSMIC database, which shares around 95% of the information contained in the Greenman/Wood dataset for the particular case of the protein kinase domain.

*Result summary and structural mutation distribution*

Finally, we wanted to assess how many of the mutations we were able to recover from the total set of mutations in the 5 studied datasets (namely, SwissProt database [31] as well as the COSMIC [32], Greenman/Wood dataset of somatic mutations [14,16], KinMutBase [8], and SAAPdb databases [33]) in order to get a view of the coverage of the existing knowledge by our method. To do so we built a non-redundant set with 1265 mutations in 317 different

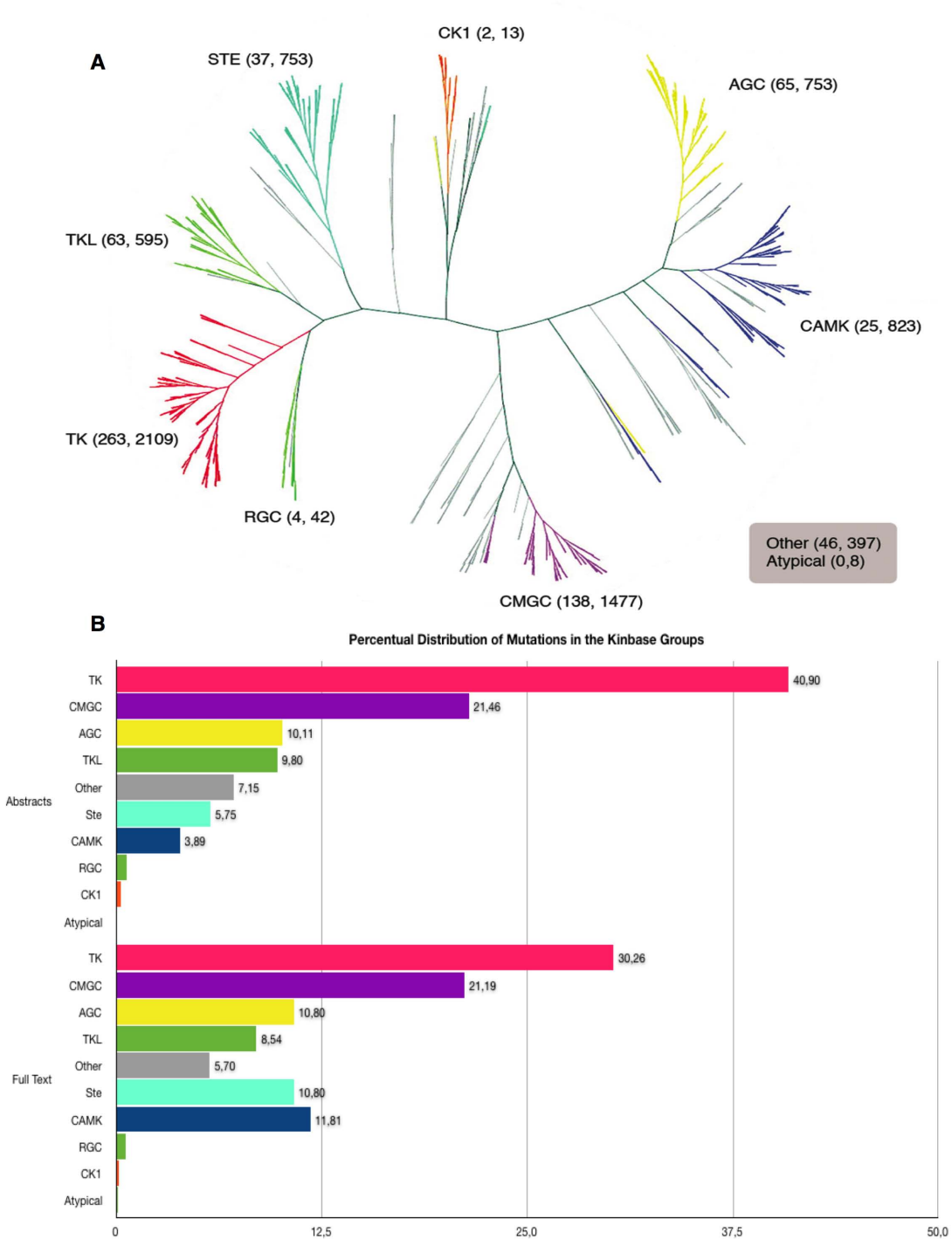



**Figure 5**

**Distribution of literature extracted mutations in the groups defined by Kinbase**. A. Number of mutations from the literature lodging in the different protein kinase domain groups in which Kinbase classifies the human kinome when the abstracts and the full text articles are taken into account respectively. B. Normalized distribution of mutations in the different protein kinase domain groups in which Kinbase classifies the human kinome when the abstracts and the full text articles are taken into account respectively

kinases. The different databases are unevenly represented, and the weight of each database is reported in the last row of Table 2, where the overlap between the different databases is assessed, under the epigraph 'All Databases'.

Out of the 1265 mutations in the combined database, 148 (11.70%) were found by the Text Mining approach when the Pubmed abstracts were scanned. By contrast 354 (27.98%) mutations were recovered when the full-text articles were taken into consideration, and 399 (31.54%) when the combined abstracts+fulltext dataset was used. The increased recall of this combined method clearly justifies the computational effort required.

Although there are mutations scattered everywhere in the kinase domain structure, a considerable mutation density is encountered close to functionally relevant parts of the protein, i.e. the ATP binding pocket, the DFG motif in the activation loop. Figure 6 shows a detailed view of the mutation density distribution within the protein kinase domain model. ATP binding Lysine 64 shows the highest density of mutations, with a total of 65 mutations, followed by residues forming the activation segment (up to 39 mutation per residue) and several residues conforming the ATP binding pocket.

***Worked example: the Epidermal Growth Factor Receptor***

The interest of the system presented here is not only that the user can gather mutations from the literature that are not reported in the databases, but also that one can get a summary of sentences mentioning those mutations that will help to assess the pathogenicity (and in the best possible scenario, the function) of the mutations newly discovered. A working example is provided here: Mutations in the EGFR.

The epidermal growth factor receptor, also known as EGFR, is a protein kinase involved in the control of cell growth and differentiation which has been reported of interest in the development of breast cancer since binding of EGF to its receptor leads to dimerization, internalization of the binary complex, induction of the tyrosine kinase activity, stimulation of cell DNA synthesis, and cell proliferation.

There are several well-known mutations reported for this protein in current state-of-the-art databases storing information on mutations (SwissProt [31], COSMIC [32], Greenman/Wood [14,16], KinMutBase [8], SAAPdb [33]) Even more, for some of them, their involvement in disease has been investigated and annotated in the corresponding databases. For instance, the somatic mutations G719S, L858R and T790M have been previously reported in relationship with lung cancer [16,36].

By contrast, our system was able to recall from the literature 32 mutation mentions that have not been reported in the dedicated databases. In order to better understand the effect of these mutations, our approach is also capable to provide context information that can be used for the interpretation of the role played by the mutations as described herein.

To provide an example, in the case of Y845F (transformed to Y869F due to the presence of a signal peptide) we were able to find the following sentences 'Furthermore, transient expression of a Y845F variant EGFR in murine fibroblasts resulted in an ablation of EGF-induced DNA synthesis to nonstimulated levels.' (PMID:10075741), 'Stably transfected B82L cells with a point mutation of the EGFR at Tyr-845 (B82L-Y845F) exhibited only basal Ras activity following exposure to Zn2+' (PMID:11983694), 'In contrast, LPA-elicited DNA synthesis and migration were augmented in cells expressing EGFR, EGFR(K721A), or EGFR(Y845F), but not EGFR(Y5F), although the PDGF responses were indistinguishable' (PUBMED 15364923). The information retrieved suggests the involvement of Tyrosine-845 from EGFR in DNA synthesis via binding to EGF.

In addition, the system also retrieves functionally neutral results that are often discarded and not stored in the databases although they contain very useful information for the contextual interpretation of the involvement of point residues in protein function 'Unexpectedly, the Y845F mutant EGFR was found to retain its full kinase activity and its ability to activate the adapter protein SHC and extracellular signal-regulated kinase ERK2 in response to EGF, demonstrating that the mitogenic pathway involving phosphorylation of Y845 is independent of ERK2-activation' (PUBMED 9990038). The structural model of this protein together with a summary of the residue and mutation information is included in additional material file 3

## Conclusion

In this paper we presented the first approach to extract human kinase mutations from both PubMed abstracts and a large collection of full text articles, comparing the obtained results to mutations that have been manually curated from the literature by annotation databases as well as data generated by genotyping studies. Automated mutation extraction can assist manual curation efforts by providing direct pointers to mutation evidence sentences for quick manual examination. The MutationFinder system was useful to detect mutation mentions from both abstracts and full text articles combined with some additional filtering of ambiguous mutation mentions. Some potential future improvements of this basic mutation extraction system could consider wrongly extracted mutation mentions resulting from mentions of sequence

ranges or the inclusion of detection of stop codons (e.g. R97X). Several strategies have been used to filter ambiguous mutation mentions and to discriminate between mutations at the level of DNA and proteins. We carried out a detailed consistency analysis of the mutations detected by means of literature mining to the content of manually curated annotations. Future steps could include a more detailed exploration of the actual reliability scoring and ranking of sequence validated mutations through the use of: (1) mutation-protein proximity analysis in full text articles, (2) species and organism source ambiguity examination and (3) analysis of the probability of finding a given mutation within the target sequence per chance, considering the actual residue composition of proteins and kinases. By using a standard machine learning approach we were able to score the level of phenotypic description based on contextual information provided for a given mutation, classifying each mutation mention as induced (artificial, generated by mutagenesis experiments) or natural variant (polymorphisms, SNPs and somatic mutations). This aspect is especially important as it connects mutation relevant information generated by different scientific domains, i.e. data generated by clinical, epidemiological and human genetics studies with molecular biology and biochemical in vitro experiments. Extraction of mutation information from multi-document collections is useful to complement different scientific discoveries and characterizations described across various papers, increasing thus efficiency in relating entries to each other and integrating multiple complementary evidences discovered by different research groups. Problems related to sequence shifts or cases of so-called sequence conflicts when comparing the numbering used by article authors to the sequences contained databases like Swiss-Prot were addressed by using various sequence validation strategies, from the basic residue look-up to the use of text derived sequence patterns. These Sequence conflicts can be the result of sequencing errors, sequence variants or isoforms that are not well characterized or even from alternative counting when considering N-terminal signal peptides [28]. To resolve such sequence conflicts is even a cumbersome task for human experts. We can recover 7,184 potential mutations on kinases in the Protein Kinase Domain from text (643 from abstracts and 6970 from full text). Information from abstracts and full text is essentially complementary, as sometimes the full text article for a mutation mentioning abstract is not available or even written in another language different from English. Although some of the extracted mutation-kinase associations might be erroneous, they still provide a very good basis for additional annotation efforts, in some cases valuable for the in depth analysis of specific proteins (as in the example shown here).

Interestingly only a very minor fraction of the mutations detected in hight throughput genotyping studies [14,16] correspond to previously identified mutations mentioned in the literature. As a considerable number of these HT generated data correspond to mutations that do not have any deleterious effect, it is understandable that they lack further careful characterization published in the literature. In general we find that 31.54% of the mutations contained in manually annotated databases can be directly recovered from papers, important for assuring the database-literature coherence. The remaining mutation records lack direct evidence about its origin in text, potentially due to (1) missing accessibility of the corresponding full text articles or additional materials (especially in case of older publications), (2) general limitations in terms of recall of mutation mention extraction methods or (3) limitations in the protein normalization and mutation to sequence associations. We estimate that, based on the proportion of natural and artificial variations described in the literature, a considerable fraction of the text mining derived mutations not contained in any of the existing kinase mutation resources might correspond to experimentally generated induced mutations. From a manual inspection of natural variant mutations we were able to differentiate between four main mutation types, some of them not considered as annotation relevant by existing databases but nevertheless important for interpreting the practical relevance of individual mutations, these include: (1) mutations with no clear association to the studied disease phenotypes, (2) mutations that are protective against some pathological condition, (3) mutations that are deleterious and that promote the pathological condition (e.g. increased disease risk). On the symmetric view 5.55% of the automatic literature annotations (23.02% from abstracts and 5.08% from full text) correspond to database confirmed entries, implying that a considerable fraction of the extracted mutations through literature mining is potential new information still to be annotated. In order to assess to which extent this new information can be trusted a human expert manual validation protocol was conducted on a randomly selected sample of 100 mutations taken from mutation mentioning abstracts (see Figure 7). We demonstrate in this work how the power of text mining combined with bioinformatics approaches can be used to discover and link information in key areas of biology, being able to result in a framework for supporting manual mutation literature curation and with the potential to adapt an analogous pipeline to other protein families going beyond the kinase/mutation analysis. Our work resulted in a collection of kinase mutation literature links (mutations, positions, sentences) derived form both full text articles and abstracts. Our work shows that extraction of mutations from full text articles is feasible and that it could be applied to the whole set of full text articles

from PubMed records in case these access to those is provide din the future.

The experiment shows that for 23% of the mutations there was a positive confirmed record in at least one of the analyzed knowledgebases (SwissProt [31], COSMIC [32], Greenman/Wood [14,16], KinMutBase [8], SAAPdb [33]), being consistent with the results previously shown for the automatic extraction pipeline. In addition, and an important added value provided by our system, 41% of the results were correct assignments between the protein and the mutation extracted by text-mining that were not reported in the knowledgebases. Finally, 8% of the mutations corresponded to orthologs having the same amino acid that the human protein at the specified position, which can be considered positive hits as well, as they essentially represent information generated for human kinases using animal models. In summary, we estimate that almost three quarters (72%) of the extracted mutations correspond to positive hits being either previously annotated mutations, correct novel mutations or mutations of close orthologs.

Interestingly, a small proportion of the records (2%) were too ambiguous even for human experts, lacking enough information even to perform manual validation.

## Materials and methods

### *Sequences of protein kinase domains using KinBase*

The KinBase resource (http://www.kinase.com/kinbase, [1]) is a repository storing the currently accepted classification of eukaryotic protein kinases, which are categorized into two main groups: 'conventional' protein kinases (ePKs) and 'atypical' protein kinases (aPKs). The ePKs form the largest group and they have been subdivided into eight groups by sequence similarity of the catalytic domains, the presence of accessory domains, and by considering different modes of regulation. The eight ePK groups defined in KinBase are: the AGC group (including cyclic-nucleotide and calcium-phospholipid-dependent kinases, ribosomal S6-phosphorylating kinases, G protein-coupled kinases and close relatives of these kinases), the CAMKs (calmodulin-regulated kinases); the CK1 group (casein kinase 1 and close relatives); the CMGC group (including cyclin-dependent kinases, mitogen-activated protein kinases, CDK-like kinases and glycogen synthase kinase); the RGC group (receptor gyanylate cyclase kinases); the STE group (MAP Kinase cascade kinases), Tyrosine kinase group (TKs); and the TKL group (Tyrosine kinase like family) which are a cluster of serine-threonine kinases resembling TKs. Another broad, miscellaneous group called 'other' is also considered for those proteins that do not fit in any of the predefined sets.

At the time of the analysis, KinBase contained 620 human protein sequences of which 518 correspond to protein kinases not considered to be pseudogenes. Although kinases described as pseudogenes are transcribed and might even have a residual or scaffolding function, kinase pseudogenes were not mapped onto Uniprot (SwissProt/Trembl) since many of them are partial transcripts or have stop codons in their sequence. Since KinBase does not directly map its entries onto Uniprot, this mapping was performed using a BlastP [37] search for each kinbase sequence against a custom database containing all entries in Uniprot annotated as human protein kinase domain. Once the mapping was performed, we were able to map 488 Kinbase identifiers to a valid Uniprot entry, 474 of them (97.13%) at sequence identity levels of at least 95%.

### *Mutation extraction from abstracts and full text articles*

The used mutation extraction pipeline has been applied to two text data sets, one consisting in the whole collection of PubMed abstracts, and the other in a set of 19,404 full text articles. The full text articles were automatically downloaded using an in house full text retrieval system that had previously been implemented. To prioritize full text articles for download, three different criteria were considered. The first selection criteria was based on information contained in the corresponding abstracts, such as mention of mutations, mention of human kinase proteins and a combination of keywords (including 'human kinase mutation'). The second selection criteria was based on extracting all the PubMed references for human kinases contained in multiple databases (e.g. SwissProt, MINT, IntAct). The third selection criteria was based on analyzing the fraction of mutation mentioning abstracts for each journal, prioritizing a set of journals (and thus their articles) for retrieving their full text articles. These journals included: the American Journal of Human Genetics, European Journal of Human Genetics, Human Genetics, Human Mutation and Human Molecular Genetics. Each of the full text articles was automatically converted into plain text using pdftotext. Both abstracts and full text articles were then preprocessed applying an in house rule-based sentence boundary detection system that we optimized for PubMed abstracts. We applied the MutationFinder system two both the full text and abstract sentence collections using a cluster of 64 Mac PPC G5 processors running Darwin.

### *Mutation disambiguation and filtering*

The performance of information extraction methods that detect mutation mentions from the literature is affected by the underlying article selection criteria used. When applied to the whole PubMed database, a fraction of extracted mutation mentions are ambiguous, and therefore can, depending on the context correspond to a range of other bio-entities, like cell lines, protein names or

**Table 2: Overlap between the different knowledgebases and the literature extracted mutations**

| | Literature derived mutations and overlap with knowledgebases | | | |
|---|---|---|---|---|
| **Knowledgebase (KB)** | **Total Mutations in KB** [weight] | **Abstract** | **Full Text** | **Combined (Abs+FT)** |
| **SwissProt – all** | 710 [56.13%] | 134 (18.87%) | 328 (46.20%) | 365 (51.41%) |
| **SwissProt – natural variant** | 459 [36.28%] | 99 (21.57%) | 196 (42.70%) | 230 (50.11%) |
| **SwissProt – mutagenesis** | 251 [19.84%] | 35 (13.94%) | 132 (52.59%) | 135 (53.78%) |
| **SAAPdb – all** | 610 [48.22%] | 65 (10.66%) | 106 (17.38%) | 125 (20.49%) |
| **SAAPdb – pathogenic deviations** | 323 [25.53%] | 64 (19.81%) | 105 (32.51%) | 123 (38.08%) |
| **SAAPdb – neutral** | 287 [22.69%] | 1 (0.35%) | 1 (0.35%) | 2 (0.70%) |
| **Greenman & Wood** | 254 [20.08%] | 4 (1.57%) | 12 (4.72%) | 13 (5.12%) |
| **Greenman & Wood – driver** | 119 [9.04%] | 3 (2.52%) | 9 (7.56%) | 9 (7.56%) |
| **Greenman & Wood – passenger** | 135 [10.67%] | 1 (0.74%) | 3 (2.22%) | 4 (2.96%) |
| **COSMIC** | 200 [15.81%] | 4 (2.00%) | 11 (5.50%) | 12 (6.00%) |
| **KinMutBase** | 83 [6.56%] | 32 (38.55%) | 32 (38.55%) | 43 (51.81%) |
| **All Databases** | 1265 | 148 (11.70%) | 354 (27.98%) | 399 (31.54%) |

clones. Only few previously published approaches did a more careful examination of wrongly extracted mutation mentions, most of these ambiguous mentions correspond to single letter mutations. Horn *et al.* compiled manually a list of exceptions to avoid mislabeling of other phrases as mutations, examining also certain terms co-mentioned in the context (e.g. cell line, tumour or cancer). For filtering single letter mentions that might correspond to mutations at the level of DNA or RNA they analyzed words surrounding the point mutation, but did not provide further details regarding this process [27]. Erdogmus and colleagues addressed DNA versus protein mutation disambiguation through a supervised learning approach based on the Naïve Bayes algorithm, they prepared a collection of 2,771 mutation mentions at the protein level and 768 at the DNA level and obtaining an accuracy of 84.7 [24].

We propose an approach for targeted mutation pattern sense disambiguation and filtering of mentions that do not correspond to protein mutations. Therefore we examined manually a large collection of mutation mentions to determine the sense inventory with respect to the context of occurrence, discriminating the main classes of false positive ambiguous mutation mentions and characterizing their semantic categories. The majority of these corresponded to one of the following three semantic types:

- Cell lines or cell types. There are several frequently mentioned cell lines that resemble mutation mentions. Among these are the human glioblastoma cell line T98G, the T-cell line M14T, the adrenocortical cell line H295R or other commonly used cell lines such as T47D or T24C.

- Taxonomic entities. Certain taxonomic names, especially bacterial strains, cloning vectors and certain animal models (e.g. mouse strains) contain words that are similar to single letter mutations. Example cases include the strains: E. coli K12S, A. viscosus T14V, P. pneumoniae R36A, A. naeslundii T14V, Mycoplasma sp. G145T or the yeast strain S288C. Also clone identifiers (e.g. W12I and W12E) or plasmids (e.g. E. coli plasmids P15A) can result in false positive mutation hits. A special case of ambiguous mutation mentions is encountered in transgenic mouse models like G93A transgenic mice. It consists in a mouse strain expressing a G93A mutant form of human SOD1 protein, but usually is mentioned as the name of the strain rather than as a reference to this particular mutation.

- Protein and gene names. Several protein names do match the patterns used to identify mutations from the literature, although some of these correspond to human proteins like S100D and S100E, a considerable fraction are viral gene names (e.g. A10L of the vaccinia virus, A11L variola virus or the poxvirus protein A52R).

We found some additional cases of wrongly tagged mutations that could be classified as drugs or compounds (e.g. the antibiotic A83586C, the immunogen A27L or the antifungal antibiotics A9145C). To determine the semantic class of a given mutation occurrence we explored the use

of knowledge-based methods relying on machine-readable dictionaries (MRDs) for sense disambiguation based on local context analysis. In order to address this disambiguation task we assumed (1) One sense per discourse, namely that within a given document the target mutation mention is consistently used as either a mutation or one of the three other semantic types previously introduced; and (2) One sense per collocation, implying that nearby co-mentioned words provide strong clues to the sense of the target mutation mention.

Three lexical resources were compiled for taxonomic entities, protein/gene names as well as cell lines. Due to limited lexical coverage of cell line information in existing biological ontologies such as the Cell Type ontology, we generated automatically a cell line dictionary through use of a named entity recognition method (ABNER, [28]) applied to mutation mentioning PubMed abstracts. This resulted in a total of 9,252 cell line names, out of which 1,124 corresponded to mentions that could potentially match mutation patterns. We incorporated from the list of cancer cell lines contained in the COSMIC database five additional names resembling mutations. This cell line dictionary was used to filter ambiguous mutation mentions (over 13,500 sentences). We also generated automatically 922 pattern templates based on multi-word cell line names, where the original word resembling a mutations is used as a slot to be filled with ambiguous mutation mentions (see table 3).

For taxonomic entities we assembled a dictionary of species names derived from the NCBI Taxonomy and used a dictionary look-up approach with these names for filtering potentially ambiguous mutations. A total of 584 taxonomic names (and their variations) contained words matching mutation regular expressions, most of them from cloning vectors and bacterial strains. Out of these we generated 128 disambiguation patterns for taxonomic mentions. A similar approach was followed for disambiguation of mutations matching protein and gene names, relying on a protein dictionary extracted from the UniProt database. The total number of protein and gene names from UniProt matching mutation mentions was 295. These were exploited for generating 29 disambiguation patterns that were manually revised to remove too general patterns, resulting finally in a set of 25 patterns.

A special case of ambiguity is encountered when distinguishing between mutations at the level of DNA, RNA and protein sequences. To enable discrimination between these different mutation types official nomenclature guidelines state that the description should be preceded by a letter indicating the type of reference sequence, p in case of protein sequences (e.g. pCys76Ala or p.C76A), g for genomic sequences, c for cDNA, m for mitochondrial sequences and r for RNA sequences [38]. Unfortunately in practice these standards are not sufficiently followed resulting commonly in ambiguity at the level of the corresponding reference sequence type, which requires a specific disambiguation strategy. This scenario is somehow similar to the distinction between gene and protein mentions, where even for human experts it is sometimes challenging to make clear decisions.

To handle the automatic distinction between DNA and protein mutations, we explored the use of different selection criteria that humans actually follow to achieve this task. We applied a hand crafted rule-based technique, with the implicit advantage that it does not require the construction of large training collections of representative sample cases for different types of DNA/protein ambiguous mutations. As contextual representation for disambiguation of mutation patterns we used: (a) implicit information from the mutation itself, i.e. mutation sequence position, (b) features derived from the local context, i.e. words enclosed in the corresponding sentence, and (c) distant content words from the whole abstract as contextual cues, i.e. other co-occurring mutations.

A useful characteristic to distinguish mutations at the DNA and protein level is actually provided by the mutation position number. The average length of sequences in UniProt is 360 amino acids, being the longest sequence 35,213 (the Titin protein from mouse). When looking at the mutation positions annotated in SwisProt, 96.76% are below 2000, 98.72% are below 3000 and 99.25% are below 4000. Therefore a basic aspect that we explored here was to filter mutations by position numbers allowing three positional cut-offs (2000, 3000 and 4000). Example cases of DNA mutations that can be successfully detected with this simple criterion are T1191C (PMID 15993850), G2950692A (PMID 15862761) and G20210A (PMID 18501222).

The local context of a given mutation mention, represented by the sentence in which it occurs can provide hints towards the mutation type. We generated two lists of terms that are associated either to mutations at the level of proteins or DNA based on manual inspection and extension of the features used by a sentence classifier trained on a small sample set of 687 DNA and protein mutation mentions. We used terms from these two lists mentioned within the mutation sentences to calculated the overlap coefficient of Lesk for scoring them as DNA or protein associated [39].

Certain distant content words co-occurring with a mutation in the whole abstract can be used as contextual cues for disambiguation. Here we explored the use of other co-

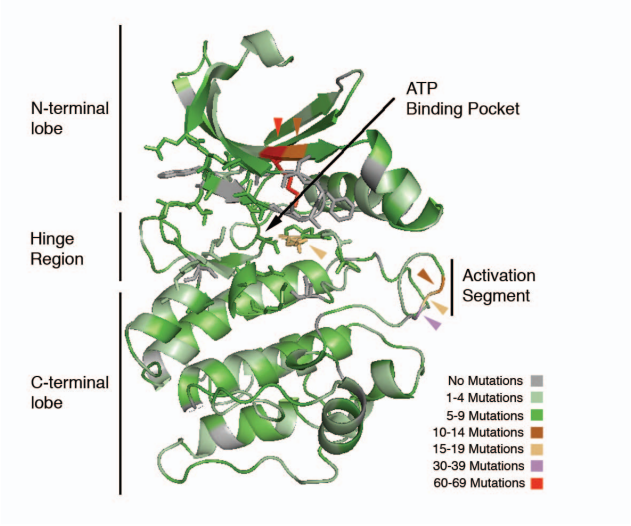



**Figure 6**
**Localization of the mutations extracted from the Pubmed abstracts within the structure of the Protein Kinase domain**. The ATP binding pocket is represented with sticks. The DFG motif (activation segment, essential for kinase function) allocates a big number of mutations. The light brown Asparagine (central part of the figure) in the inter-lobe region, more than 10 mutations. The highest density residue is Lysine 64 (red), allocating 65 mutations. This residue has been reported as essential for protein function and ATP binding. We observe that most of the mutations allocate in or near the ATP binding pocket or the activation segment and that mutations outside the binding pocket correspond generally to low mutation density residues (colored in grey and green in the kinase domain model).

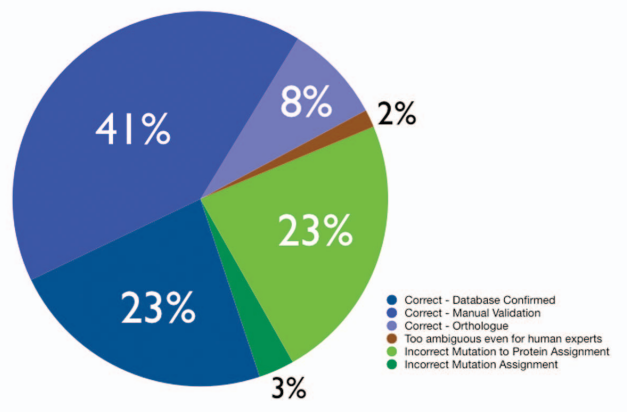



**Figure 7**
**Success estimate of the extraction pipeline by human expert manual validation**. These percentages were calculated upon a manual sampling and validation protocol conducted on 100 abstracts. Correct – Database confirmed: These are the mutations that have been found already in at least one of the analyzed databases (Uniprot, SAAPdb, COSMIC, KinMutBase or Greenman). Correct-Manual validation: This subset corresponds to the mutation-protein pairs that have been found correct after manual validation on 100 abstracts. Correct – Orthologue: This subset corresponds to the cases where mapping is confirmed by manual validation and the mutation is mapped to a non-human orthologue. Incorrect Mutation to Protein Assignment: Corresponds to the cases where both proteins share the same amino acid at the mutated position and the algorithm choses the incorrect pair. Incorrect Mutation assignment: Cases where the mutation is not properly identified. An interesting particular case are the confusion with cell lines (accounting 66% of this category) Too ambiguous even for human experts: Odd little informative cases where even human experts reading the abstracts are not able to identify to which protein the mutation corresponds to.

mentioned mutations to determine the cooperative effect for mutation disambiguation, under the assumption that if multiple mutation patterns co-occur, and all of them resemble DNA mutations, it is consequently more probable for each of them to corresponds to a DNA rather than a protein level mutation. From manual examination of the resulting hits, we determined that at least 4 distinct mutations had to be co-mentioned in a given abstract, and that at least two different mutation combinations were needed (to avoid filtering of systematic Cys to Ala-scanning mutations). An example case illustrating this idea is the PubMed record 9240741, where all the following mutations are co-occurring: T1448C, T1366G, G1604A, A1226G. Finally we also took into account the numerical relation underlying the codon triplets and their encoding for amino acids as filtering criterion for cases where for a given ambiguous mutation, another co-mentioned mutation fulfill the positional information condition: position of DNA ambiguous mutation is equal to 3 times the position of a co mentioned mutation, as illustrated for C684G and N228K in: The novel mutations include T302C (L101P), C684G (N228K), and G1063C (A354P) (PMID 9889017).

#### *Mutation phenotype level classification: natural and induced*

The classification of mutation mentions into natural variant or induced mutations was carried out using a sentence classifier approach using words co-mentioned with the mutation within the sentence. We used a SVM implementation (SVMLight, [40]) with radial basis kernel function (default parameters) which explored several feature weightings, finally using term frequency in order to avoid inconsistencies resulting from the class balance when weighting the used features. The initial feature dictionary was filtered using an in house stop word list (See additional file 4). We carried out also additional word filtering to remove numerical expressions and words with a length below 3 characters. The training set of sentences was

**Table 3: Mutation disambiguation patterns.**

| Example cases of Mutation disambiguation dictionary records and patterns | | | |
|---|---|---|---|
| **Cell lines patterns** | **Cell lines names** | **Proteins/Genes names** | **Taxonomy names** |
| human glioblastoma cell line MUTATION | breast cancer T47D cells | Met-1 serine protease | Aeromonas sp. F713E |
| MUTATION glioblastoma cell line | T98G human malignant glioma cells | S100C | Bacillus sp. G100I |
| MUTATION control cells | L5178Y lymphoblasts | Sperm surface protein P34H | Candida sp. N12C |
| MUTATION cultured cortical neurons | human cervical cancer C33A cells | R18I.1 | Synechococcus sp. D120S |
| T3 MUTATION preadipocyte clones | -BRAF (V600E) thyroid cancer cells | Protein C184L | Symbiodinium sp. H10K |

derived from mutations of proteins extracted from papers and then cross checked using the SwissProt database whether they corresponded to natural variant or mutagenesis annotations.

### *Protein and species mention detection*

For the detection of protein and organism names we used a dictionary look-up and maximum sub-string matching algorithm implemented in C and Perl. The initial gene and protein dictionary of human kinases was extracted from SwissProt and automatically extended using heuristics and rules taking into account common typographical variations encountered in gene and protein names and symbols. These covered aspects related to the use of hyphens (generating variants with hyphens, with white space and without white space), capitalization (generating variants in upper case letters and capitalized versions) and word ordering. This resulted in a human kinase protein dictionary of 2,582,220 protein name-database identifier associations. This dictionary was further manually processed based on the information content of each tagged protein mention to remove some highly ambiguous protein name variations.

### *Mutation sequence validation*

To associate co-mentioned proteins and mutations from a given article, previous efforts [19,27] often considered local text associations in terms of distances between a mutation and the nearest mentioned protein (proximity scores). These document-centric associations have clear limitations in terms of the performance, and therefore recent efforts tried to improve the underlying performance through looking up the mutation at its corresponding position within the protein sequence. In an effort to increase the recall of the method we implemented a cascade of several strategies for mutation sequence validation that included the following strategies: (1) Sliding window algorithm that searches for a pattern of mutations along the sequence instead of exact position – using the numbering given in the mutation – co-occurrences in the sequence. The algorithm iteratively scans each position in the sequence and searches for co-occurrences of the other mutations mentioned in the same abstract in positions relative to the starting one giving priority to the distance, in terms of sequence, between al the mutations in the same abstract instead of the exact positions provided. The main capability of this approach is that is able to deal with the different means in which the starting position of a protein can be defined, the most graphic case being the presence – or not – of a signal peptide but other examples can be provided (sequencing errors or discrepancies, inclusion of promoter regions, and so on. Since the finding the profile by chance is quite easy for trivial results (the easiest of them all being patterns consisting of just one mutation) a limitation in the complexity of the pattern was established, being taken into consideration only those patterns having at least 3 mutations at different sequence positions. (2) Basic mutation to sequence position mapping: looking up the wild type residue of an extracted mutation mention in the corresponding protein sequence position. (3) Alternative mutation directionality look-up: to account for errors in the automatic extraction of the mutation directionality (i.e. wild type residue with respect to mutant residue), we examined whether the mutant residue could be matched to the corresponding sequence position. (4) Pro-peptides and mature protein mutation mapping: to handle alternative residue counting when signal peptide cleavage is considered we analyzed positional wild type residue mapping for cases of proteins with N-terminal signal peptide sequences. (5) Methionine start site counting: we carried out mutation mapping taking into account as well as neglecting the N-terminal methionine.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

AV conceived the idea. AV, JMGI and MK planned the analysis. MK, JMGI and CR generated the datasets. MK

and JMGI performed the analysis. AV and MK wrote the first draft and MK and JMGI the final version. All authors read and approved the manuscript.

## Additional material

### Additional file 1

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S8-S1-S1.pdf]

### Additional file 2

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S8-S1-S2.png]

### Additional file 3

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S8-S1-S3.png]

### Additional file 4

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S8-S1-S4.txt]

## Acknowledgements

The work of the two groups in this area is funded by the ENFIN (LSHG-CT-2005-518254), MECBIO2007 (BIO2007-66855, Functions for Gene Sets) and the BIOSAPIENS (LSG-CT-2003-503265) projects and also the RD07/0067/0014 (RTIC COMBIOMED) project of the Spanish Health Ministry. We would like to thank especially Florian Leitner and also Ashish Tendulkar, Gloria Fuentes, David de Juan and Antonio Rausell as well as other members of the Valencia group for useful feedback.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 8, 2009: Proceedings of the European Conference on Computational Biology (ECCB) 2008 Workshop: Annotation, interpretation and management of mutations. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/10?issue=S8.

## References

1. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S: **The Protein Kinase Complement of the Human Genome.** *Science* 2002, **298:**1912-1934.
2. Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, Shah K, Shokat KM, Morgan DO: **Targets of the Cyclin-dependent Kinase Cdk1.** *Nature* 2003, **425:**859-864.
3. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJR, Stern DF, De Virgilio C, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF, Snyder M: **Global Analysis of Protein Phosphorylation in Yeast.** *Nature* 2005, **438:**679-684.
4. Huse M, Kuriyan J: **The conformational plasticity of protein kinases.** *Cell* 2002, **109(3):**275-82.
5. Burgess AW: **EGFR family: structure physiology signalling and therapeutic targets.** *Growth Factors* 2008, **26(5):**263-74.
6. Yamada S, Shiro Y: **Structural basis of the signal transduction in the two-component system.** *Adv Exp Med Biol* 2008, **631:**22-39.
7. Sanz P: **AMP-activated protein kinase: structure and regulation.** *Curr Protein Pept Sci* 2008, **9(5):**478-92.
8. Ortutay C, Väliaho J, Stenberg K, Vihinen M: **KinMutBase: a registry of disease-causing mutations in protein kinase domains.** *Hum Mutat* 2005, **25(5):**435-42.
9. Richardson CJ, Gao Q, Mitsopoulous C, Zvelebil M, Pearl LH, Pearl FMG: **MoKCa Database-mutations of Kinases in Cancer.** *Nucleic Acids Res* 2009, **37:**D824-D831.
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-42.
11. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2008:D281-8.
12. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA: **The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution.** *Nucleic Acids Res* 2007:D291-7.
13. Shchemelinin I, Sefc L, Necas E: **Protein kinases, their function and implication in cancer and other diseases.** *Folia Biol (Praha)* 2006, **52(3):**81-100.
14. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber TD, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JKV, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PVK, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B: **The genomic landscapes of human breast and colorectal cancers.** *Science* 2007, **318(5853):**1108-13.
15. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JKV, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE: **The consensus coding sequences of human breast and colorectal cancers.** *Science* 2006, **314(5797):**268-74.
16. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446(7132):**153-8.
17. Baudot A, Real F, Izarzugaza J, Valencia A: **From cancer genomes to cancer models: bridging the gaps.** *EMBO Rep* 2009.
18. Santamaría D, Barrière C, Cerqueira A, Hunt S, Tardy C, Newton K, Cáceres JF, Dubus P, Malumbres M, Barbacid M: **Cdk1 is sufficient to drive the mammalian cell cycle.** *Nature* 2007, **448(7155):**811-5.
19. Rebholz-Schuhmann D, Marcel S, Albert S, Tolle R, Casari G, Kirsch H: **Automatic extraction of mutations from Medline and cross-validation with OMIM.** *Nucl Acids Res* 2004, **32:**135-142.
20. Lee LC, Horn F, Cohen FE: **Automatic Extraction of Protein Point Mutations Using a Graph Bigram Association.** *PLoS Comput Biol* 2007, **3:**e16-e16.
21. Baker CJO, Witte R: **Mutation Mining – A Prospector's Tale.** *Information Systems Frontiers (ISF)* 2006, **8:**47-57.
22. Witte R, Baker CJO: **Towards A Systematic Evaluation of Protein Mutation Extraction Systems.** *J Bioinform Comput Biol* 2007, **5(6):**1339-1359.

23. Caporaso JG, Baumgartner WA, Randolph DA, Cohen KB, Hunter L: **MutationFinder: a High-performance System for Extracting Point Mutation Mentions from text.** *Bioinformatics* 2007, **23:**1862-1865.
24. Erdogmus M, Sezerman OU: **Application of Automatic Mutation-gene pair Extraction to Diseases.** *J Bioinform Comput Biol* 2007, **5:**1261-1275.
25. McDonald R, Scott Winters R, Ankuda CK, Murphy JA, Rogers AE, Pereira F, Greenblatt MS, White PS: **An Automated Procedure to Identify Biomedical Articles that Contain Cancer-associated gene Variants.** *Hum Mutat* 2006, **27:**957-964.
26. McDonald RT, Winters RS, Mandel M, Jin Y, White PS, Pereira F: **An Entity Tagger for Recognizing Acquired Genomic Variations in Cancer Literature.** *Bioinformatics* 2004, **20:**3249-3251.
27. Horn F, Lau AL, Cohen FE: **Automated Extraction of Mutation data from the Literature: Application of MuteXt to G Protein-coupled Receptors and Nuclear Hormone Receptors.** *Bioinformatics* 2004, **20:**557-568.
28. Yip YL, Lachenal N, Pillet V, Veuthey AL: **Retrieving Mutation-specific Information for Human Proteins in UniProt/Swiss-Prot Knowledgebase.** *J Bioinform Comput Biol* 2007, **5:**1215-1231.
29. Kanagasabai R, Choo KH, Ranganathan S, Baker CJO: **A Workflow for Mutation Extraction and Structure Annotation.** *J Bioinform Comput Biol* 2007, **5:**1319-1337.
30. Yip YL, Famiglietti M, Gos A, Duek PD, David FPA, Gateau A, Bairoch A: **Annotating Single Amino acid Polymorphisms in the UniProt/Swiss-Prot Knowledgebase.** *Hum Mutat* 2008, **29:**361-366.
31. Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch A: **The Swiss-Prot variant page and the ModSNP database: A resource for sequence and structure information on human protein variants.** *Human Mutation* 2004, **23(5):**464-470.
32. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, Wooster R: **The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website.** *Br J Cancer* 2004, **91(2):**355-8.
33. Hurst J, McMillan L, Porter C, Allen J, Fakorede A, Martin A: **The SAAPdb web resource: A large-scale structural analysis of mutant proteins.** *Hum Mutat* 2009.
34. Caporaso JG, Baumgartner WA, Randolph DA, Cohen KB, Hunter L: **Rapid Pattern Development for Concept Recognition Systems: Application to Point Mutations.** *J Bioinform Comput Biol* 2007, **5:**1233-1259.
35. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, Sun C, Liu Hh, Torres R, Krauthammer M, Lau WW, Liu H, Hsu CN, Schuemie M, Cohen KB, Hirschman L: **Overview of BioCreative II gene Normalization.** *Genome Biol* 2008, **9(Suppl 2):**S3-S3.
36. Tam IYS, Chung LP, Suen WS, Wang E, Wong MCM, Ho KK, Lam WK, Chiu SW, Girard L, Minna JD, Gazdar AF, Wong MP: **Distinct epidermal growth factor receptor and KRAS mutation patterns in non-small cell lung cancer patients with different tobacco exposure and clinicopathologic features.** *Clin Cancer Res* 2006, **12(5):**1647-53.
37. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17):**3389-402.
38. den Dunnen JT, Antonarakis SE: **Mutation Nomenclature.** *Curr Protoc Hum Genet* 2003, **Chapter 7(Unit 7.13):**.
39. Lesk M: **Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone.** *Proceedings of SIGDOC-86: 5th International Conference on Systems Documentation* 1986:24-26.
40. Joachims T: *Learning to Classify Text using Support Vector Machines* 2002 [http://www.cs.cornell.edu/People/tj/]. *Kluwer*
41. Saunders RE, Perkins SJ: **CoagMDB: a Database Analysis of Missense Mutations Within four Conserved Domains in five Vitamin K-dependent Coagulation Serine Proteases Using a Text-mining tool.** *Hum Mutat* 2008, **29:**333-344.
42. Furlong LI, Dach H, Hofmann-Apitius M, Sanz F: **OSIRISv1.2: a Named Entity Recognition System for Sequence Variants of Genes in Biomedical Literature.** *BMC Bioinformatics* 2008, **9:**84-84.

EMBO reports

review

# From cancer genomes to cancer models: bridging the gaps

*Anaïs Baudot[1+], Francisco X. Real[2,3], José M. G. Izarzugaza[1] & Alfonso Valencia[1]*

[1]Structural Biology and Biocomputing Programme, and [2]Molecular Pathology Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain, and [3]Universitat Pompeu Fabra, Barcelona, Spain

**Cancer genome projects are now being expanded in an attempt to provide complete landscapes of the mutations that exist in tumours. Although the importance of cataloguing genome variations is well recognized, there are obvious difficulties in bridging the gaps between high-throughput resequencing information and the molecular mechanisms of cancer evolution. Here, we describe the current status of the high-throughput genomic technologies, and the current limitations of the associated computational analysis and experimental validation of cancer genetic variants. We emphasize how the current cancer-evolution models will be influenced by the high-throughput approaches, in particular through efforts devoted to monitoring tumour progression, and how, in turn, the integration of data and models will be translated into mechanistic knowledge and clinical applications.**

Keywords: cancer bioinformatics; cancer genome; cancer models; driver/passenger; genetic variants

EMBO *reports* (2009) **10,** 359–366. doi:10.1038/embor.2009.46

## Introduction

Cancers result from the accumulation of genetic changes (Vogelstein & Kinzler, 2004), and the identification of gene variants involved in tumour development and progression has been a central goal of cancer research for years (Sidebar A). Projects such as the Human Cancer Genome Project, The Cancer Genome Atlas and the International Cancer Genome Consortium aim to decipher the spectrum of genetic variants in different cancer types. The goals of these high-throughput resequencing (HTR) studies are fourfold: to identify genetic changes associated with tumour phenotypes; to discover molecular biomarkers that might be used for early detection, more accurate diagnosis or prognosis; to determine the molecular events of tumorigenesis; and, ultimately, to use this knowledge to develop strategies for targeted therapy (Chin & Gray, 2008; Wood *et al*, 2007).

[1]Structural Biology and Biocomputing Programme, and [2]Molecular Pathology Programme, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro 3, E-28029 Madrid, Spain
[3]Department de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Edifici PRBB, Dr. Aiguader 88, E-08003 Barcelona, Spain
[+]Corresponding author. Tel: +34 917 328 000; Fax: +34 912 246 976;
E-mail: abaudot@cnio.es

**Submitted 13 November 2008; accepted 23 February 2009; published online 20 March 2009**

However, there is an intense debate about the extent to which large-scale variation data will help us to understand the molecular mechanisms of tumour evolution. It is fair to say that, so far, the first genome-wide cancer HTR projects have had limited impact on molecular cancer research. These studies are rarely quoted as a starting point for further experiments (supplementary Table S1), although it is clear that more time is needed to translate gene discovery into mechanistic understanding. Technical, cultural and scientific issues can be responsible for the gap between genomic data and outcomes in terms of the molecular understanding of tumorigenesis. In the first place, the current methods for the organization of genomic data are evolving along with sequencing developments and constitute a real handicap for the use of the information. Second, high-throughput technologies unavoidably generate noise; the computational and statistical methods used to filter out genomic data—on which the reliability of the observations provided to the community ultimately depend—are not exempt from complications. Third, the core of the scientific challenge lies in the difficulty of linking genomic data to the molecular processes that underlie cancer evolution, as discussed in the final section of this review. It is therefore not surprising that cancer genome initiatives have generated substantial criticism, as many biologists are used to (and favour) more targeted approaches (Chng *et al*, 2007; Loeb & Bielas, 2007; Strauss, 2007).

## The mutational landscape of tumours

Many types of genetic variant contribute to cancer: small structural changes (such as point mutations or small insertions), major structural rearrangements (such as translocations), numerical changes and epigenetic changes (supplementary Table S2). Alterations in the control of aneuploidy could also have a role (Duesberg, 2007). Mutations can occur spontaneously in cancer cells—through cytosine deamination, for example—after exposure to carcinogens or as the result of a mutator phenotype caused by mutations in polymerases and/or in mismatch-repair genes, which can lead to chromosomal instability (Loeb *et al*, 2008). In principle, all genes that harbour modifications are candidate cancer genes.

Genetic variants can be transmitted through the germline or can arise through somatic mutation. Germline variants are present in all the cells of an individual and contribute to inherited cancer susceptibility. One particular case of germline variants are the single-nucleotide polymorphisms (SNPs), the most common genetic variants, which are, by definition, present in at least 1% of the population (Collins

**Sidebar A** | In need of answers

(i) How can we define a cancer gene and how many cancer genes exist?
(ii) How can the functional effects of mutations in cancer cells be predicted?
(iii) How can cancer genes and their associated functional roles be precisely assessed by detailed biological research?
(iv) How can the gene variants that are causally involved in tumour development or progression be identified among all the gene variants in a tumour?
(v) How can the gap between the genetic variants observed in cancers and the current models of cancer evolution be bridged?
(vi) How can the experimental analysis of gene variants involved in cancer be accelerated?

*et al*, 1998; The International HapMap Consortium, 2003). Germline variants are typically identified through resequencing, and their involvement in cancer is shown using linkage or association studies (supplementary Table S2). Somatic mutations arise in the genomes of dividing cells and, in fact, all adult organisms are probably mosaics of somatically mutated cells. Somatic changes are typically identified through the resequencing of candidate genes, by analysing chromosomal rearrangements, or by quantifying losses or gains in gene-copy numbers using a range of techniques—such as microsatellite analysis or the quantitative polymerase chain reaction (qPCR; supplementary Table S2). Evidence for epigenetic silencing and downregulation of expression provides further support for the identification of tumour-suppressor genes, whereas increased expression can provide evidence for oncogene identification. The experimental validation of the biochemical and/or biological effects of a given alteration is often considered as proof of mechanistic involvement. In this context, RNA interference provides an additional method to study the involvement of gene variants in tumorigenesis. It has, for example, been recently used to validate mouse tumour-suppressor candidates (Zender *et al*, 2008).

The detection of point mutations has generally been carried out with small-scale sequencing from one to a few genes; >25,000 mutations identified in the well-known tumour-suppressor *TP53* (Soussi *et al*, 2000) have been collected using this approach.

In the past decade, technical advances have provided the opportunity to use high-throughput methods for the identification of candidate cancer genes. Functional genomics approaches—such as microarray or methylation studies—have also been used, as well as association analyses and, more recently, tumour HTR screenings to determine the genes responsible for the initiation and progression of cancer (supplementary Table S2).

## Large-scale resequencing studies

HTR studies can detect point mutations and short insertions or deletions (Bardelli *et al*, 2003); the introduction of 'next-generation sequencing' technologies (Mardis, 2008) has not only produced massive amounts of data, but also allows the quantitative identification of individual gene variants and the detection of abnormal transcripts (Campbell *et al*, 2008). So far, HTR studies have followed two approaches, focusing either on genes or on tumour types (Table 1). In the first approach, a subset of genes—such as those that encode protein kinases (Greenman *et al*, 2007)—is sequenced in a relatively large number of samples. This approach allows the identification of genes that are mutated at low frequencies, but also requires an *a priori* selection of genes. The second approach analyses the coding sequences of whole genomes in a smaller number of tumour samples, and has been applied to colon and breast tumours (Sjöblom *et al*, 2006; Wood *et al*, 2007), pancreas adenocarcinomas (Jones *et al*, 2008) and glioblastoma (Parsons *et al*, 2008). This approach allows for the identification of the most-frequently mutated genes (Table 1). One such HTR study screened 518 protein kinases in 26 primary lung neoplasms and seven cell lines, and identified 188 mutations in 141 genes (Davies *et al*, 2005).

**Table 1** | Catalogue of main recent high-throughput cancer genomic studies and initiatives

| First author | Publication date | Genes | Tumours | Screen sizes | PMID |
|---|---|---|---|---|---|
| Bardelli | 2003 | Tyrosine kinase | Colon | 138 genes, 35 samples, a subset in 147 additional samples | 12738854 |
| Wang | 2004 | Tyrosine phosphatase | Colon | 87 genes, 18 samples, a subset in 157 additional samples | 15155950 |
| Stephens | 2005 | Kinase | Breast | 518 genes, 25 samples, a subset in 56 additional samples | 15908952 |
| Davies | 2005 | Kinase | Lung | 518 genes, 33 samples, a subset in 56 additional samples | 16140923 |
| Sjöblom | 2006 | All | Breast and colon | 13,023 genes, 22 samples, a subset in 48 additional samples | 16959974 |
| Greenman | 2007 | Kinase | 210 human cancers | 518 genes in 210 samples | 17344846 |
| Wood | 2007 | All | Breast and colon | 18,191 genes, 22 samples, a subset in 48 additional samples | 17932254 |
| Loriaux | 2008 | Tyrosine kinase | Acute myeloid leukaemia | 85 genes, 188 samples | 18252861 |
| Tomasson | 2008 | Tyrosine kinase | Acute myeloid leukaemia | 26 genes, 94 samples, a subset in 94 additional samples | 18270328 |
| Brown | 2008 | Tyrosine kinase | Chronic lymphocytic leukaemia | 70 genes, 95 samples | 18754031 |
| Jones | 2008 | All | Pancreas | 20,661 genes, 24 samples | 18772397 |
| Parsons | 2008 | All | Glioblastoma | 20,661 genes, 22 samples, a subset in 83 additional samples | 18772396 |
| CGARN | 2008 | 601 genes | Glioblastoma | 601 genes, 91 samples | 18772890 |
| Ding | 2008 | 623 genes | Lung | 623 genes, 188 samples | 18948947 |

CGARN, Cancer Genome Atlas Research Network; PMID, PubMed identifier.

**Table 2** | Main cancer-specific and non-specific repositories that contain information about cancer-associated mutations

| Acronym | Full name* | Category | URL | Reference |
|---|---|---|---|---|
| COSMIC | Catalogue of Somatic Mutations in Cancers | Mutations | http://www.sanger.ac.uk/genetics/CGP/cosmic | Forbes *et al*, 2008 |
| CGC | Cancer Gene Census | Cancer genes | http://www.sanger.ac.uk/genetics/CGP | Futreal *et al*, 2004 |
| OMIM | Online Mendelian Inheritance in Man | Disease-related genes | http://www.ncbi.nlm.nih.gov/omim | Hamosh *et al*, 2005 |
| Ensembl | – | Polymorphisms | http://www.ensembl.org | Flicek *et al*, 2008 |
| dbSNP | Single Nucleotide Polymorphism Database | Polymorphisms | http://www.ncbi.nlm.nih.gov/SNP | Sherry *et al*, 2001 |

*For information on additional repositories, please consult supplementary Table 3.

Regardless of the strategy used, these studies produce an overwhelming amount of information. The results are usually provided as raw tables in the supplementary material of a given publication and the main outcomes are briefly summarized in the published text. A number of databases aim to compile this type of information, such as Catalogue of Somatic Mutations in Cancer (COSMIC), which lists >60,000 mutations (Forbes *et al*, 2008), and the Cancer Gene Census (CGC), which—as of October 2008—included data for 380 cancer genes (Futreal *et al*, 2004); other repositories also include cancer-related information (Table 2; supplementary Table S3). However, although these genomic data are of great biological value, they are, in general, not sufficiently linked to additional information on gene annotation and regulation, or on molecular interactions and pathways, or to the clinical data about tumour and tissue types. In analysing this panorama, one realizes that the available infrastructure for organizing cancer genome information is still in its infancy and certainly lags behind the capacity of the current massive experimental approaches.

## Drivers and passengers

Like all high-throughput approaches, HTR generates noise that is difficult to distinguish from real biological signals. This noise can be technical, coming directly from sequencing technologies or from limitations in tumour-cell collection; all methods are sensitive to the presence of the normal allele, either in tumour cells or in contaminating normal cells. Gene variants that correspond to SNPs are ideally pinpointed by sequencing both tumour and normal tissues from the same patient, or by checking polymorphism databases. However, the most important source of problems is the presence of numerous mutations that are clearly detectable but do not have a direct role in cancer. In fact, only a handful of gene mutations that have been identified in HTR studies are likely to be biologically meaningful. To distinguish these mutations from the background mutation noise is a difficult task.

Mutations can be classified as 'drivers' or 'passengers' depending on their involvement in cancer development and progression. This metaphor was probably used for the first time in 1964, during a keynote lecture by Sir Christopher Andrewes, in which he referred to the role of viruses in either causing cancer (drivers) or being merely passengers in infected cells (Andrewes, 1964). Today, the term driver is used to denote mutations and/or genes that are positively selected and contribute to tumour development or progression, whereas the term passenger is used to designate cancer-neutral variations that are retained during the evolution of the cancerous cells.

Single mutations can be responsible for the development and progression of a cancer (Fig 1A). Historically, analyses have focused on mutations that can affect protein function. These mutations are thought to be mainly non-synonymous (missense, nonsense or frameshift), in contrast to synonymous (silent) mutations. In this regard, the first oncogene identified—H-*ras*—was found to have a non-synonymous substitution in codon 12 that introduces an alanine in the position of a glycine, thereby blocking its GTPase function and producing a protein able to transform cells (Reddy *et al*, 1982). However, this corresponds to a 'protein-centric' view of biology; one should remember that non-synonymous mutations might not always alter protein function (owing to amino-acid plasticity) and, importantly, that there is strong evidence showing that 'silent' mutations can be biologically relevant—for example, through the modulation of splicing (Cartegni *et al*, 2002)—although it is difficult to assess their effects and even more difficult to predict them. Additionally, we have to keep in mind that 98% of the genome is intergenic. In this respect, it is currently impossible to interpret the consequences of mutations in non-coding DNA regions, with the exception of some favourable cases in splice sites or promoters.

The current way of thinking assumes that only a small fraction of the non-synonymous mutations actually cause tumours. Historically, the identification of mutations has been followed by functional analyses to evaluate their pathogenic potential. For example, a screening of the gene encoding the tyrosine kinase FLT3 (FL cytokine receptor) identified nine non-synonymous mutations (Fröhling *et al*, 2007), four of which allow the growth of cultured cells independently of the presence or absence of growth factors. In general, only a small range of biological assays is used to assess pathogenicity, exploring a limited spectrum of the potential biological effects of candidate mutations and often being unable to detect small functional changes (Chin & Gray, 2008). When a direct effect on cell proliferation or the generation of apoptosis is not detected, other experiments are seldom used unless the functional annotations point directly to a crucial biological role—as is the case for proteases and kinases. It is a formidable challenge to scale up these experiments to validate the results of genome-wide HTRs, and, therefore, *in silico* methods are a suitable alternative. Typical computational methods are based on the assumption that somatic mutations considered as 'drivers' would have to affect protein function markedly (Torkamani & Schork, 2007). Sequence and protein domain conservation, as well as protein structure, are used to determine the crucial positions in a given protein and to predict the causative effects of mutations (Fig 1A). The same parameters are also applied to predict the possible pathogenicity of SNPs (supplementary information online).

Bioinformatic predictions based on sequence analysis—made by the SIFT (Ng & Henikoff, 2003) and PMut (Ferrer-Costa *et al*, 2005)

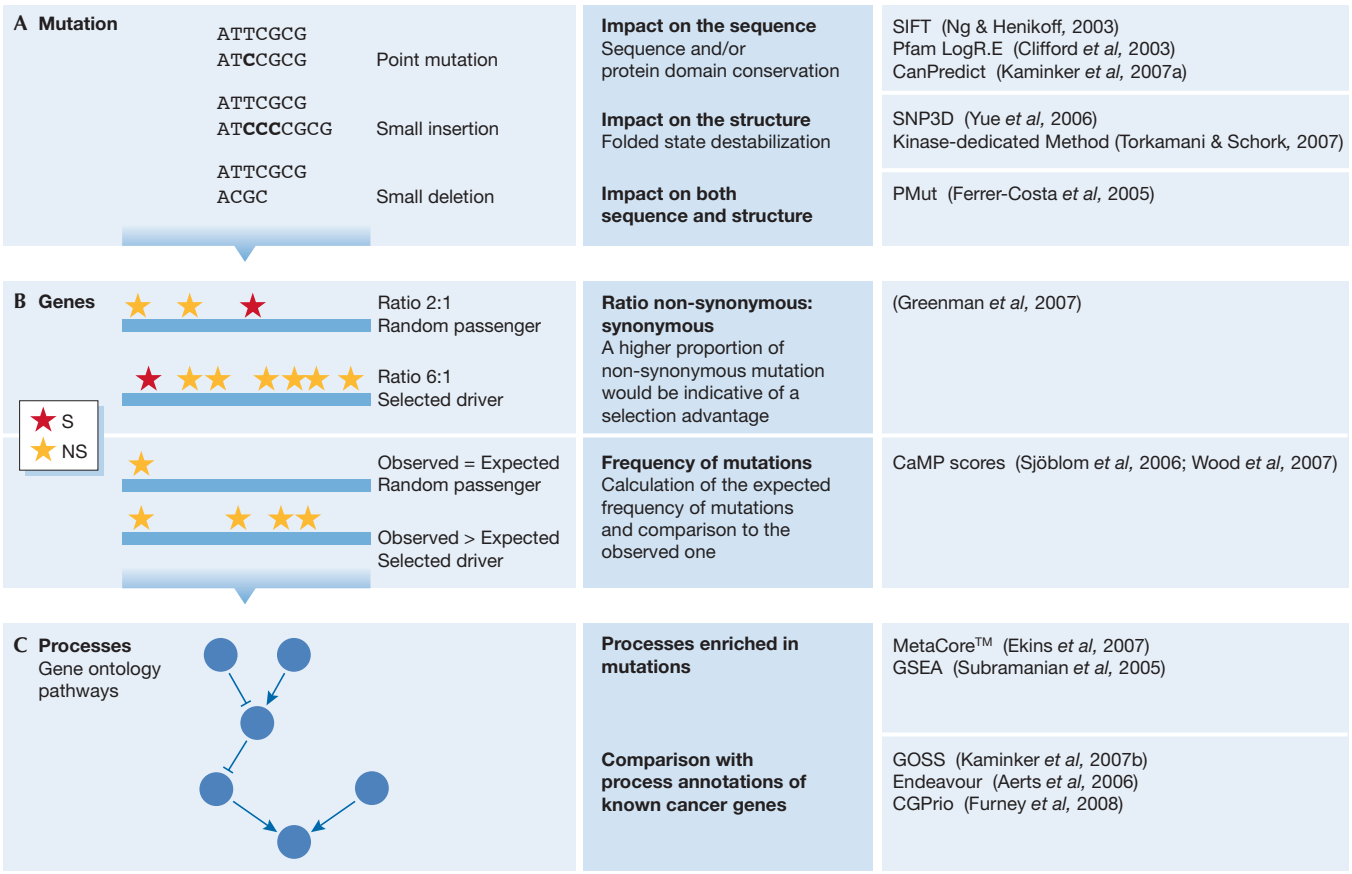



**Fig 1** | Driver or passenger? Multilevel strategies used to classify mutations and genes as either 'drivers' or 'passengers' at the level of (**A**) mutations, (**B**) genes or (**C**) processes. NS, non-synonymous; S, synonymous.

programs—and experimental results have been compared for nine *FLT3* mutations (Fröhling *et al*, 2007). The consensus bioinformatic predictions failed to identify two mutations that were experimentally shown to affect function, whereas one predicted driver mutation was not found to have a functional effect. By contrast, all four of the mutations predicted to be passengers *in silico* were confirmed by the functional analysis. The availability of more case studies will allow a better assessment of the predictive capacity of computational tools.

It is important to remember that there is a great difference between demonstrating that a mutation alters the function of a protein and claiming that it has a pathogenic involvement in cancer. The effects of candidate mutations in the development of cancer are probably highly context dependent and the assessment of their biological significance in the context of human cancer needs to be largely extrapolated.

At a second level of analysis, the overall frequency of mutations in a given gene can help to detect a positive selection that would support its involvement as a driver for oncogenesis (Fig 1B). The usual assumption is that positive selection is exerted mainly on non-synonymous mutations. The genetic code provides a random ratio of approximately two non-synonymous mutations for each synonymous one (2:1); higher ratios are interpreted as evidence of positive selection and competitive advantage. In reality, more complex models—borrowed from the field of molecular evolution—are applied, which can also take into account the types of mutation (transition and transversion) or the neighbouring sequence (for example, whether a C-to-T transition occurs in a CpG island; a more detailed explanation of the models used can be found in the supplementary information online). Another set of methods calculates differences between the observed and the expected frequencies of non-synonymous mutations. If a gene contains—in all sequenced tumours—more mutations than would have been expected to occur by chance, these have been positively selected during the process of tumorigenesis and, therefore, confer an advantage in this process (supplementary information online).

An obvious limitation of these approaches to the identification of cancer genes is the need to sequence many samples. Without enough observations, the less-frequently mutated genes would not meet the statistical thresholds. In fact, they would be indistinguishable from unselected passengers, although they can be revealed by functional assays (Fröhling *et al*, 2007). Furthermore, these statistical techniques do not provide information about the specific alleles—point mutations—involved in cancer evolution. A perhaps less obvious—albeit not less important—limitation of current studies is that they usually consider mutations individually, without modelling epistatic interactions. In a few cases, the importance of the combination of otherwise neutral (passenger) mutations has been shown (Chen *et al*, 2008). Epistatic effects (Moore, 2005), which are not commonly considered in cancer genome studies, might be even more important when taking genetic background into consideration, either alone or together with somatic mutations.

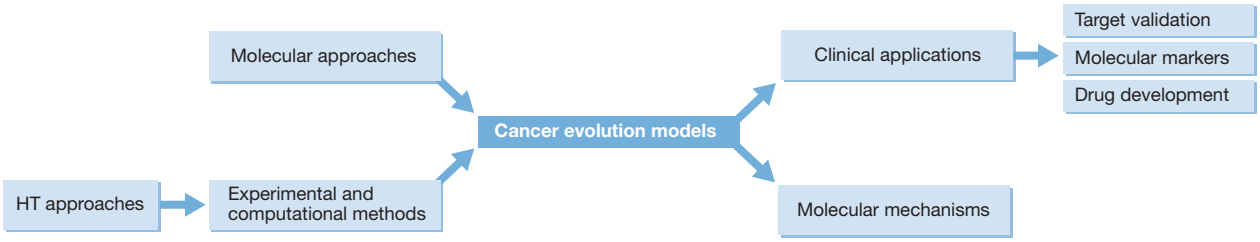



**Fig 2** | Modelling cancer evolution. Using the available technologies, the modelling of cancer evolution should provide insights into its development and progression.

Among known mutations, a large proportion occurs in a few genes, such as *TP53* or *K-RAS* (Forbes *et al*, 2008; Soussi *et al*, 2000). Hence, cancer genomes are composed of a handful of frequently mutated genes and a much larger number of infrequently mutated genes. The number of genes that are mutated in cancers, although large, possibly reflects alterations in a relatively small number of signalling pathways (Wood *et al*, 2007; Fig 1C). Indeed, many recent HTR studies provide an interpretation of their results in terms of alterations in 'core pathways' (Jones *et al*, 2008; Parsons *et al*, 2008). This point is crucial—particularly from a therapeutic point of view—because designing strategies to target proteins individually is different from targeting a well-defined pathway (Check Hayden, 2008). Additionally, cancer genes might share other structural or functional properties (Furney *et al*, 2006), such as good evolutionary conservation or a role in essential cellular processes such as the cell cycle or DNA repair. The analyses based on previous knowledge of known pathways and functions can be useful for the interpretation of genome-wide results. However, to obtain new insights into the oncogenic process, it is important to avoid the constant re-identification of the same genes for which significant functional information is already available.

## Cancer-evolution models and cancer genomics

Molecular biologists have been working for the past 20 years to determine the molecular mechanisms of cancer evolution. Modelling cancer evolution is more than an academic exercise as it has profound implications on the detection of early recurrence and in the choice of adjuvant therapy, among other aspects (Fig 2). To be useful in this context, large-scale genomic studies would have to complement these efforts, and help to improve our understanding of tumour development and progression.

Historically, cancer research has been dominated by the 'clonal evolution' model of tumour development and progression (Fig 3, blue arrows). This model postulates that tumour cells acquire specific genetic changes, leading to clonal expansion. These changes are selected in competition with other tumour cells through a Darwinian process, and those that confer a selective advantage become fixed, thereby allowing a phylogenetic tracing of the history of the evolving cell populations. In this model, it is generally considered that benign lesions are precursors of malignant tumours, genomically stable tumours precede genomically unstable ones and metastases are the ultimate step in tumour evolution. More recently, it has become evident that some experimental results do not fit this model. For example, some metastases of breast cancers bear little genetic resemblance to the primary tumour (Schmidt-Kittler *et al*, 2003). Moreover, a recent study showed that a small proportion of normal mouse mammary epithelial cells injected intravenously can survive at distant sites and eventually develop into tumours (Podsypanina *et al*, 2008). These observations have led to the proposal of the 'parallel evolution' model (Gray, 2003; Yokota & Kohno, 2004), in which cells that generate metastases are separated relatively early from the primary tumour and evolve independently (Fig 3, red arrows). This model is reinforced by data from gene-expression profiles that are predictive of metastases in certain primary breast tumours (Bernards & Weinberg, 2002).

The differences between both models have important consequences for the interpretation and clinical use of the knowledge about cancer-associated mutations. In the framework of the 'clonal evolution' model, the significance of a mutation detected in a metastasis—in the absence of information about its presence in the primary tumour or in pre-neoplastic lesions—is unclear. In the context of the 'parallel evolution' model, targeting this mutation for therapy would have no effect on the growth of the primary tumour, which could evolve to metastasis through other mutations. Given the complexity of cancer and the diversity of its phenotypic presentation, it is unlikely that a single paradigm will universally account for cancer development and progression; the different models might be complementary rather than exclusive, at least when considering cancer globally rather than at the individual level. It seems possible that these two models might explain different, albeit concurrent, biological processes (Fig 3, green arrows). This integration is also reinforced by evidence that metastases can act as repositories from which additional systemic tumour-cell seedings can take place (Nguyen & Massagué, 2007).

HTR studies are usually performed using DNA from cell lines, xenografts or large and advanced tumours. This bias in sample selection, owing to the fact that earlier stage tumours are underrepresented, is to some extent also present in low-scale studies. During the advanced stages of tumorigenesis, all the mutations necessary for cancer development and progression are already present; as is commonly assumed, "such tumours contain all the mutations found in the early stage tumours, but the converse is not true" (Wood *et al*, 2007). The information derived from HTR studies is therefore intrinsically far from providing information on other stages of cancer evolution, and hence does not contribute to our understanding of the development and evolution of cancer. Furthermore, the identification of causative cancer genes and mutations—based on the methods adapted from evolutionary biology described above—tends to be too general to give specific information at the level of resolution required by the current cancer-evolution models. Hence, we continually have to revisit our understanding of the contribution of genetic variants based only on the study of snapshots in tumour evolution, which do not provide sufficient insight to elucidate the true relevance of these genes in the tumorigenic process.

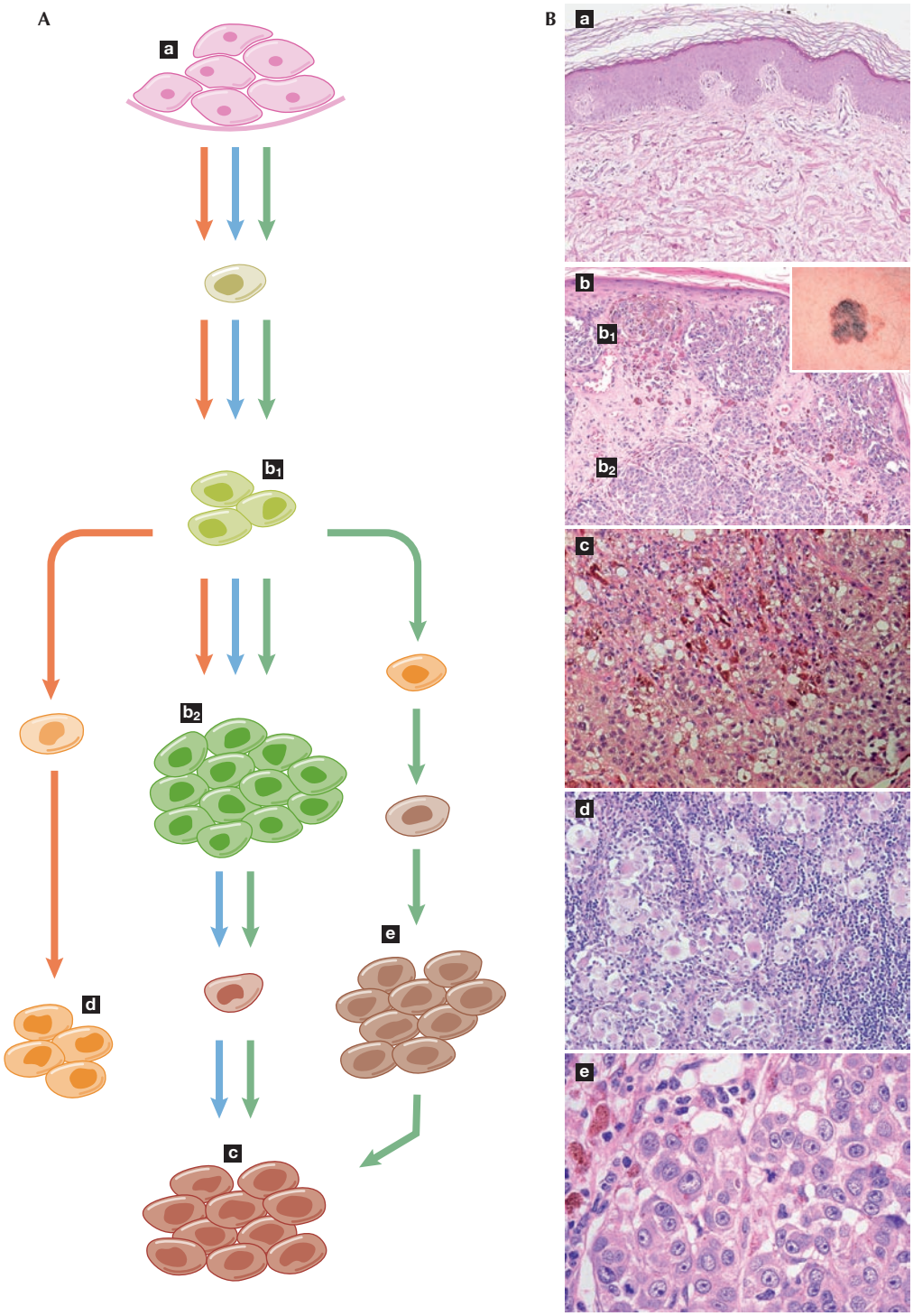



**Fig 3** | Models of cancer evolution. (**A**) The 'clonal selection model' (blue arrows) is the prevailing view to explain the successive steps of mutation and selection from normal tissue to primary tumour and metastasis. However, metastasis-generating cells can emerge relatively early in the tumorigenic process and 'seed' distant tissues, thereby evolving in parallel with the primary tumour and delineating the 'parallel evolution' model (red arrows). Finally, these two models can occur simultaneously and metastatic deposits can act as sites from which additional metastases can be generated, therefore leading to an integrated model of cancer evolution (green arrows). (**B**) Microphotographs provide a histological snapshot of normal skin tissue (**a**), primary tumour (superficial $\mathbf{b}_1$ and deep $\mathbf{b}_2$, macroscopic appearance inset in **b**), subcutaneous metastasis (**c**), metastasis in the lymph node (**d**) and metastasis in the lung (**e**), and are shown in correspondence with the cancer-evolution models. This melanoma—which originates from the transformation of pigmented skin cells—provides a visual example of the modelling paradigms, illustrating the gap between ideal models and actual observations.

In order to attain more insight into the contribution of the different models to cancer development, and to validate more precisely the significance of the genetic and genomic changes found in advanced tumours, one would need to obtain information about specific genes and mutations at different stages of tumour evolution. Knowing the time of appearance of a given mutation would allow for a better estimate of its contribution to the fitness of cancer cells, which is essential to distinguish between the various evolution models. This involves several difficulties and only a few metastasis-specific alterations have been identified (Nguyen & Massagué, 2007). First, it is conceivable that individual genetic alterations, or a given genetic programme, could render a stage-specific advantage to tumours and be either neutral or deleterious at later stages, as is the case for the epithelial–mesenchymal transition programme, which is activated in the invasive front of tumours but might be repressed in metastases (Thiery & Sleeman, 2006). Additionally, changes in the tumour microenvironment—either locally or at metastatic sites—might impose different selection pressures on genetic changes and thereby modulate the influence of the individual mutations. Furthermore, the effects of genetic alterations might be incremental rather than qualitative, thereby allowing for epistatic interactions, which are often not considered when modelling molecular pathogenesis. Improvements in technology and more focused research on early-stage tumours are needed to fill these gaps. For such applications, the lack of sensitivity for detecting a given mutation in a low proportion of alleles is a major technical concern when standard sequencing technology is used. However, this limitation might be overcome with ultra-sequencing technologies (Gupta, 2008). These technologies—which are already able to detect rare subclones with a sensitivity as low as 1 in 5,000 copies—would be relevant to track the subpopulations of cells that are responsible for initiating the genetic lesions, for drug resistance or for metastasis (Campbell *et al*, 2008).

Integrative approaches would be a solution to overcome the limitations specific to both genomic and functional methods. The findings obtained using diverse high-throughput genomic techniques—such as gene mutation, copy-number variation, expression analyses and epigenetic changes—have recently been combined for glioblastoma and pancreatic ductal adenocarcinoma (Cancer Genome Atlas Research Network, 2008; Jones *et al*, 2008; Parsons *et al*, 2008). The gathering of independent evidence supported the causative implication of genes in tumours, for example, by showing that a subset of the genes recurrently found in copy-number-alteration regions has an expression pattern that correlates with copy number (Cancer Genome Atlas Research Network, 2008). In practice, the interpretation of heterogeneous high-throughput information is still a formidable challenge, and multidimensional analyses of data coming from high-throughput studies still face the problems of data standardization, database annotations and normalization of phenotypic descriptions.

The combination of all these efforts should have an impact on the development of improved strategies for early detection, improved tumour subclassification, a more rational selection of therapy and more accurate prognostication, all of which represent important aspects of patient management.

## Conclusion

Cancer genome studies—including the inevitably associated computational analyses—have the potential to predict which genes and mutations contribute to tumour development (known as driver genes or mutations) on a large scale. However, despite the enormous capacity of the experimental resequencing methodologies and the expected improvements therein, limitations still exist. Indeed, the reliable detection of less-frequent mutations is still arduous, and it is difficult to obtain a sufficiently systematic mutation analysis that will allow conclusions to be drawn about the prevalence and distribution of mutations according to tumour stage. Furthermore, mutation analysis can by itself provide only statistical information on potential associations with cancer and not direct causative information, and it is a major challenge in molecular terms to go from genomic information to data interpretation. For example, the classification of mutations as drivers or passengers depends on the analysis of the possible functional consequences of these mutations, which is a technology that is not free from limitations and, in addition, does not provide a complete picture of the actual implication of the mutations in the development of cancer. In other words, the future challenge will be to support—or to refute—the current cancer models with high-throughput experimental methods within a reasonable time scale at an affordable cost. This would involve both the descriptive large-scale genomic analysis of pre-neoplastic lesions and early cancers, and the functional analysis of genetic variants: a combined effort that is crucial to translate genomic knowledge into molecular pathophysiology and patient management.

We must note that many of the difficulties in the application of high-throughput variation approaches are similar to those found in the study of other complex diseases. Cancer is particularly challenging—and therefore attractive—as this is the field in which the largest amount of molecular information is available, the diversity of phenotypes and pathologies is more notable, and the complex evolution of disease at the cellular and/or tissue level has been most directly addressed. These are all good reasons to believe that the symbiosis of high-throughput technologies, molecular and cellular mechanistic models, and new experimental systems and models will be effective first in cancer research.

**Supplementary information** is available at *EMBO reports* online (http://www.emboreports.org)

ACKNOWLEDGEMENTS

We thank M. Soengas and A. Toll for providing microphotographs. This work is supported by Instituto de Salud Carlos III (ISCIII) grant COMBIOMED (RD07/0067/0014), BIO2007-66855, the Spanish Ministry of Education and Science, and European Union grant LSHG-CT-2003-503265 (BioSapiens). Work in the F.X.R. laboratory is supported by grant SAF2007-60860 and CONSOLIDER INGENIO BioCancer from Ministerio de Ciencia e Innovación. A.B. is supported by the Juan de la Cierva fellowship.

REFERENCES

Aerts S *et al* (2006) Gene prioritization through genomic data fusion. *Nat Biotechnol* **24:** 537–544

Andrewes C (1964) Tumour-viruses and virus-tumours. *BMJ* **1:** 653–658

Bardelli A *et al* (2003) Mutational analysis of the tyrosine kinome in colorectal cancers. *Science* **300:** 949

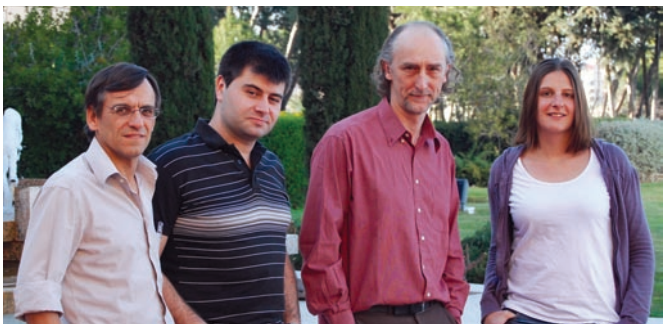
Bernards R, Weinberg RA (2002) A progression puzzle. *Nature* **418:** 823

Brown JR, Levine RL, Thompson C, Basile G, Gilliland DG, Freedman AS (2008) Systematic genomic screen for tyrosine kinase mutations in CLL. *Leukemia* **22:** 1966–1969

Campbell PJ *et al* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40:** 722–729

Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455:** 1061–1068

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3:** 285–298
Check Hayden E (2008) Cancer complexity slows quest for cure. *Nature* **455:** 148
Chen Z, Feng J, Saldivar J, Gu D, Bockholt A, Sommer SS (2008) EGFR somatic doublets in lung cancer are frequent and generally arise from a pair of driver mutations uncommonly seen as singlet mutations: one-third of doublets occur at five pairs of amino acids. *Oncogene* **27:** 4336–4343
Chin L, Gray JW (2008) Translating insights from the cancer genome into clinical practice. *Nature* **452:** 553–563
Chng WJ *et al* (2007) Limits to the Human Cancer Genome Project? *Science* **315:** 762; author reply 764–765
Clifford RJ, Edmonson MN, Nguyen C, Buetow KH (2004) Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics* **20:** 1006–1014
Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* **8:** 1229–1231
Davies H *et al* (2005) Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res* **65:** 7591–7595
Ding L *et al* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455:** 1069–1075
Duesberg P (2007) Chromosomal chaos and cancer. *Sci Am* **296:** 52–59
Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T (2007) Pathway mapping tools for analysis of high content data. *Methods Mol Biol* **356:** 319–350
Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* **21:** 3176–3178
Flicek P *et al* (2008) Ensembl 2008. *Nucleic Acids Res* **36:** D707–D714
Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **10:** 10.11
Fröhling S *et al* (2007) Identification of driver and passenger mutations of FLT3 by high-throughput DNA sequence analysis and functional assessment of candidate alleles. *Cancer Cell* **12:** 501–513
Furney SJ, Higgins DG, Ouzounis CA, López-Bigas N (2006) Structural and functional properties of genes involved in human cancer. *BMC Genomics* **7:** 3
Furney SJ, Calvo B, Larrañaga P, Lozano JA, Lopez-Bigas N (2008) Prioritization of candidate cancer genes—an aid to oncogenomic studies. *Nucleic Acids Res* **36:** e115
Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR (2004) A census of human cancer genes. *Nat Rev Cancer* **4:** 177–183
Gray JW (2003) Evidence emerges for early metastasis and parallel evolution of primary and metastatic tumors. *Cancer Cell* **4:** 4–6
Greenman C *et al* (2007) Patterns of somatic mutation in human cancer genomes. *Nature* **446:** 153–158
Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol* **26:** 602–611
Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33:** D514–517
Jones S *et al* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321:** 1801–1806
Kaminker JS, Zhang Y, Watanabe C, Zhang Z (2007a) CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res* **35:** W595–W598
Kaminker JS *et al* (2007b) Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res* **67:** 465–473
Loeb LA, Bielas JH (2007) Limits to the Human Cancer Genome Project? *Science* **315:** 762; author reply 764–765
Loeb LA, Bielas JH, Beckman RA (2008) Cancers exhibit a mutator phenotype: clinical implications. *Cancer Res* **68:** 3551–3557
Loriaux MM *et al* (2008) High-throughput sequence analysis of the tyrosine kinome in acute myeloid leukemia. *Blood* **111:** 4788–4796
Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* **24:** 133–141
Moore JH (2005) A global view of epistasis. *Nat Genet* **37:** 13–14
Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31:** 3812–3814
Nguyen DX, Massagué J (2007) Genetic determinants of cancer metastasis. *Nat Rev Genet* **8:** 341–352
Parsons DW *et al* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* **321:** 1807–1812
Podsypanina K, Du YN, Jechlinger M, Beverly LJ, Hambardzumyan D, Varmus H (2008) Seeding and propagation of untransformed mouse mammary cells in the lung. *Science* **321:** 1841–1844
Reddy EP, Reynolds RK, Santos E, Barbacid M (1982) A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300:** 149–152
Schmidt-Kittler O *et al* (2003) From latent disseminated cells to overt metastasis: genetic analysis of systemic breast cancer progression. *Proc Natl Acad Sci USA* **100:** 7737–7742
Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29:** 308–311
Sjöblom T *et al* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* **314:** 268–274
Soussi T, Dehouche K, Béroud C (2000) p53 website and analysis of p53 gene mutations in human cancer: forging a link between epidemiology and carcinogenesis. *Hum Mutat* **15:** 105–113
Stephens P *et al* (2005) A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat Genet* **37:** 590–592
Strauss BS (2007) Limits to the Human Cancer Genome Project? *Science* **315:** 762–764; author reply 764–765
Subramanian A *et al* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102:** 15545–15550
The International HapMap Consortium (2003) The International HapMap Project. *Nature* **426:** 789–796
Thiery JP, Sleeman JP (2006) Complex networks orchestrate epithelial–mesenchymal transitions. *Nat Rev Mol Cell Biol* **7:** 131–142
Tomasson MH *et al* (2008) Somatic mutations and germline sequence variants in the expressed tyrosine kinase genes of patients with *de novo* acute myeloid leukemia. *Blood* **111:** 4797–4808
Torkamani A, Schork NJ (2007) Accurate prediction of deleterious protein kinase polymorphisms. *Bioinformatics* **23:** 2918–2925
Vogelstein B, Kinzler KW (2004) Cancer genes and the pathways they control. *Nat Med* **10:** 789–799
Wang Z *et al* (2004) Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science* **304:** 1164–1166
Wood LD *et al* (2007) The genomic landscapes of human breast and colorectal cancers. *Science* **318:** 1108–1113
Yokota J, Kohno T (2004) Molecular footprints of human lung cancer progression. *Cancer Sci* **95:** 197–204
Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* **7:** 166
Zender L *et al* (2008) An oncogenomics-based *in vivo* RNAi screen identifies tumor suppressors in liver cancer. *Cell* **135:** 852–864



***Francisco X. Real, José M. G. Izarzugaza, Alfonso Valencia & Anaïs Baudot***

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

# Cancer-associated mutations are preferentially distributed in protein kinase functional sites

Jose M. G. Izarzugaza,[1] Oliver C. Redfern,[2] Christine A. Orengo,[2] and Alfonso Valencia[1]*

[1] Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro 3, Madrid E28029, Spain

[2] Biochemistry and Molecular Biology Department, University College London, University of London, London WC1E 6BT, United Kingdom

ABSTRACT

**Protein kinases are a superfamily involved in many crucial cellular processes, including signal transmission and regulation of cell cycle. As a consequence of this role, kinases have been reported to be associated with many types of cancer and are considered as potential therapeutic targets. We analyzed the distribution of pathogenic somatic point mutations (drivers) in the protein kinase superfamily with respect to their location in the protein, such as in structural, evolutionary, and functionally relevant regions. We find these driver mutations are more clearly associated with key protein features than other somatic mutations (passengers) that have not been directly linked to tumor progression. This observation fits well with the expected implication of the alterations in protein kinase function in cancer pathogenicity. To explain the relevance of the detected association of cancer driver mutations at the molecular level in the human kinome, we compare these with genetically inherited mutations (SNPs). We find that the subset of nonsynonymous SNPs that are associated to disease, but sufficiently mild to the point of being widespread in the population, tend to avoid those key protein regions, where they could be more detrimental for protein function. This tendency contrasts with the one detected for cancer associated-driver-mutations, which seems to be more directly implicated in the alteration of protein function. The detailed analysis of protein kinase groups and a number of relevant examples, confirm the relation between cancer associated-driver-mutations and key regions for protein kinase structure and function.**

Proteins 2009; 00:000–000.


**Key words: cancer; kinase; kinome; functional; somatic mutation; point mutation; driver mutation; passenger mutation; polymorphism; SNP; single nucleotide polymorphism.**

## INTRODUCTION

Point mutations are important events in the evolution of proteins and therefore organisms. There are multiple factors that influence whether a given mutation is accepted and this is often based on how specific protein characteristics are affected. For example, a mutation might alter functional residues at catalytic sites, residues that determine the specific binding of effectors, or general biophysical properties such as protein folding and stability, or protection against misfolding. Equally, a mutation can have little or no effect on the function or stability of a protein.

The most common biologically relevant mutations are single nucleotide polymorphisms (SNPs), which account for about 90% of sequence polymorphisms in humans[1] at an overall frequency of about one per 1000 bases in DNA[2] (i.e., in translated regions) or non-coding according to their genomic location. Coding SNPs are further subclassified according to whether they alter the composition of the translated protein (Nonsynonymous SNPs, nsSNPs), either through amino acid substitution or by the generation of truncation mutations. By contrast, synonymous SNPs (also referred to as silent or sSNPs) are those that do not affect the amino acid sequence of the protein product. Not all synonymous SNPs are completely neutral since they may still affect the expression of gene products or protein translation by introducing alterations into the regulatory region, interfering with splice sites or impinging on any other regulatory mechanism.[3,4] Equally, it is also the case that not all nsSNPs are associated with pathological diseases, since some changes are by nature milder than others, and diseases commonly involve complex sets of alterations.

The role of point mutations is especially pertinent in the development of cancer where large numbers of somatic mutations accumulate in the cell during tumorigenesis. These somatic

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: EMBRACE; Grant number: LSG-CT-2004-512092; Grant sponsor: BIOSAPIENS; Grant number: LSG-CT-2003-503265.

*Correspondence to: Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro 3, Madrid E28029, Spain.
E-mail: valencia@cnio.es

Received 29 January 2009; Revised 2 June 2009; Accepted 4 June 2009
Published online 19 June 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22512

changes to a cell occur either as a result of random errors during replication or might be directly induced by exposure to carcinogens.[5] Evidence suggests that only a small number of these mutations actually contribute to tumorigenesis, by conferring these malignant cells with a competitive growth advantage, and often described as "driver mutations." By contrast, the vast majority of somatic mutations that accumulate in tumors during clonal expansion do not necessarily confer growth advantage and are generally biologically neutral ("passenger mutations").[6,7]

Despite the importance of mutations in the development of malignant tissues, only around 1% of all human genes are known to contribute to cancer in this way and the genes that are most frequently involved are those that encode members of the protein kinase superfamily.[8] These enzymes are therefore obvious pharmaceutical targets for cancer therapies as they are involved in a wide range of tumorigenic activities, such as immune evasion, proliferation, anti-apoptotic activity, metastasis, and angiogenesis.[9]

Protein kinases are the most ubiquitous superfamily of signaling molecules in human cells, accounting for ~2% of the proteins encoded by the human genome.[10] They can be further divided into subfamilies that share significant similarity both at the sequence and structural level, which is understandable as all kinases transfer the terminal phosphate of ATP to serine, threonine, or tyrosine residues of a target protein. Furthermore, empirical studies also suggest a common catalytic mechanism whereby ATP and an active site divalent cation are bound in identical manners, and phospho-transfer is carried out by a shared set of amino acids. Despite this experiments in yeast models, Refs. 11, 12 have shown the protein kinase superfamily to be highly promiscuous as a whole, phosphorylating a large number of different protein substrates, although individual subfamilies display remarkable specificity.[10] This inconsistency suggests that kinases have a domain committed to the general function of catalysis, whereas another region (or even regions) may confer substrate specificity to the enzyme, without altering the general kinase folding, interfering with ATP binding or the general reaction mechanism.

The aim of this study was to analyze the distribution of different types of point mutations observed across the catalytic domain (namely the PK domain) of the protein kinase superfamily (including those specifically associated with cancer) with respect to functionally and structurally important regions. We compared the location of synonymous (sSNPs), nonsynonymous (nsSNPs), somatic "driver," and "passenger" mutations, with respect to three main categories: (a) evolutionary conservation at the primary sequence level; (b) structurally important regions; and (c) functional regions, including substrate-binding sites and the positions relevant to the specific recognition of modulators and effectors. Indeed, a related set of characteristics, based on both sequence and structure, is being used by state-of-the-art predictors of pathogenicity to determine whether a mutation can affect protein function and therefore be, potentially, associated to disease. This is still a hot topic as shown by the considerable number of publications on the subject during the last few years. Some of the more relevant ones are described in Refs. 13–24.

## RESULTS

Known germline SNPs were obtained from dbSNP for the protein kinase superfamily, along with somatic mutations[6,7] that had been further classified as "driver" (i.e., disease-associated) and "passenger" (i.e., those point mutations not thought to be pathogenic). We present the results of mapping these different types of mutations onto a representative structural model from the protein kinase superfamily [Fig. 1] and analyzing their distribution relative to evolutionary conserved positions and known functional regions (buried, functional, conserved, etc).

In the presented results, to assess the significance of the proximity of different sets of mutations to specific areas of the protein, we use the Xd measure.[25] The reason for choosing this weighted measure of distance distributions is to give priority to the differences in regions closer to the studied regions (for example, binding sites) over differences in the distribution of residues in positions far away from the regions of interest. Xd has been used as a standard measure of the difference between distributions of predicted residues in the context of the CASP challenge.[26,27] Full details are given in the "Materials and Methods" section.

### Cancer mutations in relation to sequence conserved regions

We first examined the distances of the mutated residues from sequence conserved positions in the different protein kinases [Fig. 2(b) and Supporting Information Fig. S1]. Figure 2(a) shows that driver mutations are significantly closer than passenger ones to these conserved positions, both when defined in terms of sequence identity or similarity. This difference was also visible in terms of Xd values[25] (Xdpassenger-Xddriver [21bins] = −1.23, Supporting Information Table S1). As expected, the comparison between nonsynonymous and synonymous SNPs revealed that nsSNPs tend to be further away from conserved sequences, particularly for the more strict definition of conservation using sequence identity (XdsSNPs-XdnsSNPs [21bins] = 0.89, Supporting Information Table S1a). In addition, the same differences were visible when conservation was evaluated in terms of amino acid variability using AL2CO,[28] that is driver mutations are closer to conserved positions than passenger ones (Xdpassenger-Xddriver = −0.88, Supporting Information

Fig. S2 and Supporting Information Table S1c). The same trend was observed for the synonymous SNPs that are closer to conserved positions (Xd = −0.59) than nonsynonymous ones (Xd = 0.17), respectively.

### Cancer mutations in relation to structurally conserved regions

The distance between our different classes of mutations and structurally conserved positions was also quantified using the Xd values (Supporting Information Table S2) but produced no difference between the distributions of driver/passenger mutations (ΔXd = −0.08) or of nsSNPs and sSNPs (ΔXd = 0.11) when a global score was used to assess structural conservation. However, using a local score—where only the vectors to residues within a radius of 10 Å were taken into account—there were clear differences in the distributions. When comparing nsSNPs with sSNPs, we found that nonsynonymous SNPs tend to be further away from structurally conserved residues than synonymous ones, with a difference in Xd values of 1.43. Likewise, passenger mutations tend to be closer to structurally conserved residues than driver mutations (difference of the Xd values = 1.57)

### Cancer mutations and solvent accessibility

We also analyzed the distance of the mutated positions to "buried" residues (Supporting Information Fig. S3), defined as those with a Naccess' relative residue solvent accessibility score of less than 16%, and found no statistical difference between driver and passenger mutations (Xd difference = −0.01, Supporting Information Table S3). There was, however, a difference between nonsynonymous and synonymous SNPs (Xd Difference = 2.26), where synonymous SNPs tended to be closer to buried positions than nonsynonymous SNPs (Supporting Information Fig. S4). This is consistent with the idea that residues in the buried core of proteins tend to be conserved and that these positions tend to change less.

### Cancer mutations in relation to known functional regions

We analyzed the distribution of the known mutations with respect to two definitions of a functionally important region: those directly related with ATP binding; and those related to binding sites specific for the union of effectors and modulators of kinase activity.

#### *Cancer mutations versus functional binding sites*

The functionally active region of protein kinases includes the ATP binding site, the peptide-substrate-binding sites, and the catalytic loop that is implicated in phosphate transfer. The ATP binding pocket has three parts as follows:

1. a region of hydrophobic residues clustered around the adenosine of ATP;
2. an area around the Gamma-phosphate of ATP and the divalent cation (the catalytic site) which is primarily enclosed by charged residues;
3. a region in the large lobe composed of both hydrophobic and polar residues below the ATP that stabilizes this region, and may play a role in mediating substrate interactions.

The key residues that are responsible for the positioning of ATP and stabilizing the active conformation in the catalytic mechanism are (according to Ref. 29, see Fig. 1—These positions correspond to the generated structural model):

i. Lysine 74 that interacts with the alpha and beta phosphates of ATP and stabilizes it;
ii. a nearby Glutamic acid (E96) that forms a salt bridge with lysine 74 and increases the stability of this network;
iii. Aspartate 171 that serves as the catalytic base that initiates phospho-transfer by deprotonating the acceptor serine, threonine, or tyrosine;
iv. Asparagine 176 which interacts with a secondary divalent cation, thereby positioning the gamma-phosphate of ATP; and finally,
v. Aspartate 190 that chelates the primary divalent cation, indirectly positioning ATP at the same time.

In the peptide binding site, the conservation of the substrate-binding groove is particularly important (Fig. 1), located between the catalytic loop, the P+1 loop (activation segment), helix D, helix F, helix G, and helix H (each secondary structure element is labeled with a letter in alphabetic order corresponding to its position within the protein chain, see Fig. 1). We have used the set of binding residues extracted from the FireDB database as an operational definition of the kinase binding site,[30] which includes 32 residues directly contacting the ATP in the binding pocket [Fig. 3(a)] and which covers the five highly conserved residues mentioned above (K74, E96, E171, N176, and E190).

When the distance distribution of point mutations was examined [Fig. 3(b,c)], the driver mutations appear to be closer than passenger mutations to the ATP binding pocket, both as defined by FireDB[30] and Knight et al.[29] This tendency was stronger for the residues in direct contact with the substrate identified by FireDB [Fig. 3(a), and Supporting Information Tables S4a and S4b]. As expected, nonsynonymous SNPs tend to be located further away from the binding sites than synonymous SNPs, with a difference in Xd values of 0.97 when using the FireDB definition of the binding site and of 1.02 for the subset described by Knight et al.

#### *Cancer mutations in relation to functional regions involved in binding specificity*

Residues differentially conserved in the various groups of protein kinases were identified for each of the eight

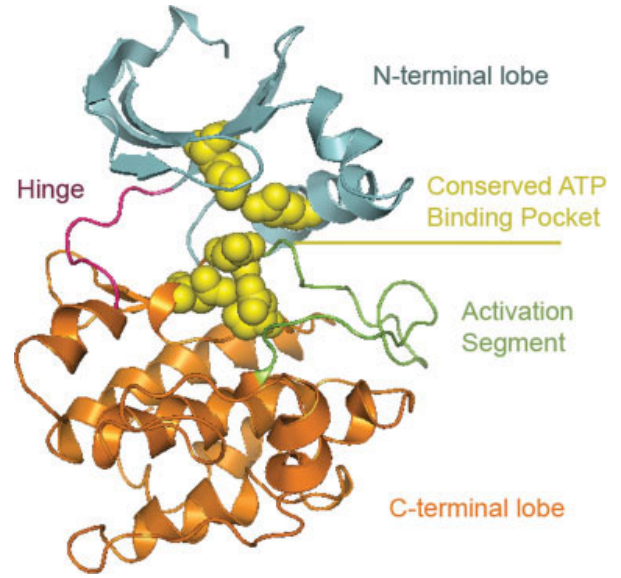



**Figure 1**

Our model structure of human protein kinase, based on MAP3K1, shows the basic two-lobe kinase fold, with the N- and C-terminal (cyan and orange, respectively) lobes joined by a hinge region (magenta). The main recognition for the substrate protein is through interaction with the activation segment (green), a region in the C-terminal lobe. ATP binds at a site between the two lobes, where five highly conserved residues*: K74, E96, D171, N176, and D190 (yellow, numbers corresponding to positions in the generated structural model) guide the positioning of the molecule. By contrast, the substrate-binding groove is located between the catalytic loop, the P+1 loop (activation segment), helix D, helix F, helix G, and helix H.

groups in which Kinbase categorizes the human kinome[10] using the appropriate bioinformatics methods (tree-determinant residues,[31,32] S3det Rausell et al., in preparation]). We identified 35 residue positions as containing sufficient information to differentiate between the groups—i.e., residues that are conserved in specific kinase groups but that are not conserved across all kinases. The most statistically significant Tree-Determinants were distributed among the kinase groups as follows: four in the AGC, four in CK1, eight CMGC, and nine STE. All tree-determinant positions were mapped onto the representative structure (Supporting Information Fig. S5) and visual inspection revealed that many of the Tree-Determinants were located near the ATP/ pocket of the protein. Other sets of tree-determinants were located in regions that are known to interact with other protein partners and that participate in intermolecular signaling events.[33] We used the position of the tree-determinant residues as a proxy for functionally important regions in protein kinases, particularly those related with the specific functions of each of the groups.

The distribution of point mutations with respect to the position of tree-determinants was determined (Supporting Information Fig. S6) and the corresponding Xd values were compared (Supporting Information Table S5). Driver mutations showed a clear tendency to be closer to tree-determinant positions than passenger mutations. Likewise, synonymous SNPs also tend to be closer to the tree-determinants, even if this tendency was not very strong, reflecting the capacity to tolerate such substitutions in functionally important regions.

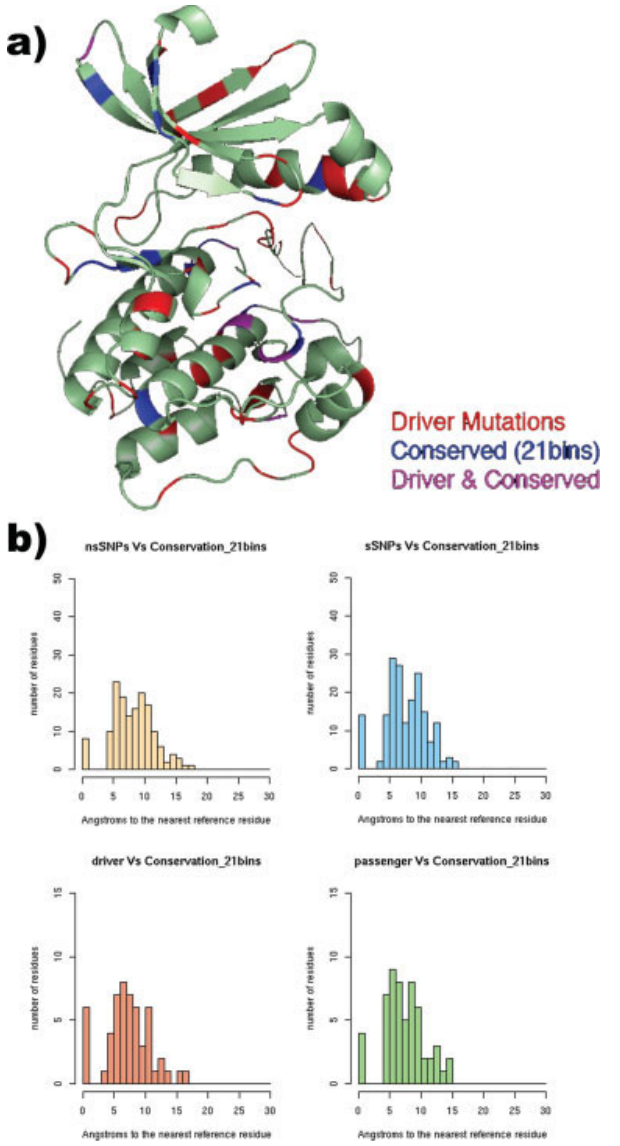



**Figure 2**

(**a**) Distribution of driver mutations compared with positions predicted to be conserved in terms of Shannon's Entropy in the context of identity (21bins). (**b**) Distribution of distances between point mutations (driver mutations, passenger mutations, nonsynonymous SNPs, and synonymous SNPs) and conserved residues in an "identity" scenario where 21 bins are taken into account, one for each aminoacid and another extra bin for positions with gaps.

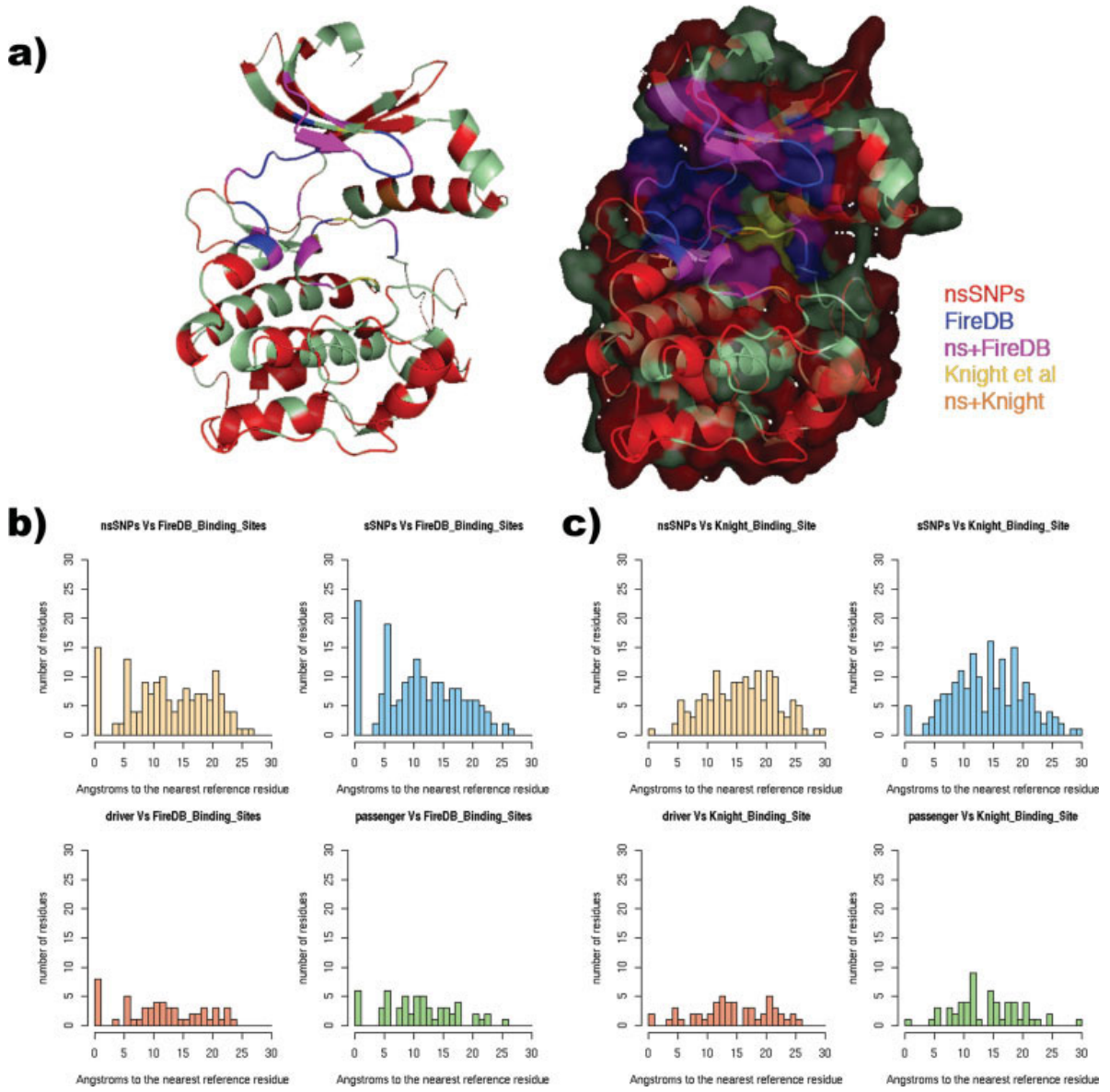



**Figure 3**

(**a**) Distribution of nsSNPs and compared with FireDB predicted residues as part of the ATP binding pocket. The figure also illustrates the five highly conserved residues: K74, E96, D171, N176, and D190 (yellow) that guide the positioning of the ATP molecule. (**b**) Distribution of distances between point mutations (driver mutations, passenger mutations, nonsynonymous SNPs, and synonymous SNPs) with respect to the ATP binding pocket, as described in FireDB (Van de Waals radii +0.5 A). (**c**) Distribution of distances between point mutations (driver mutations, passenger mutations, nonsynonymous SNPs, and synonymous SNPs) with respect to the ATP binding pocket, as described by Knight et al.

## Analysis of cancer mutations in specific protein kinase groups

To complete our sequence analysis of the distribution of point mutations, we carried out a systematic comparison of the distribution of residues in the most important protein kinase families, selecting those cases for which enough sequence/structural information was available. For this analysis, we collected all the SNPs in the proteins belonging to each group provided by Kinbase[10] in a single representative set for each protein kinase superfamily. This approach makes the number of SNPs (Supporting Information Table S6) comparable to those of the cancer related mutations specific to each group.

The comparison between nonsynonymous and all SNPs shows that driver mutations in general tend to cluster closer to tree-determinants than nonsynonymous SNPs.

This is evident in Supporting Information Tables S6, S7a, S7b, and S8, in which we have represented the differences between Xd values corresponding to each group of SNPs and mutations, subdivided into all, passenger and driver mutations, with respect to several protein features.

One interesting case study is the comparison between driver mutations and nonsynonymous SNPs, since they might also distort protein function and are potentially implicated in the development of disease. In most of the groups, driver mutations are closer to the conserved sequence regions than nonsynonymous SNPs, irrespective of the method used to quantify conservation (Supporting Information Tables S6 and S8). A similar trend was observed in terms of functional sites (FireDB and Knight datasets), the only exceptions residing in the tyrosine kinase like (TKL) and "other" groups when the binding sites were taken from FireDB. In terms of accessibility to the solvent, driver mutations were closer than nonsynonymous SNPs to buried positions, with some exceptions in groups such as AGC, Calmoldulin regulated kinases (CAMK), and MAPK cascade kinases (STE). These tendencies highlight the relationship between driver mutations and the positions that are potentially important for the correct functioning of the protein, such as conserved positions, scaffolding positions in the hydrophobic core of the protein, and residues that are important for the activity of the protein (e.g., those in the ATP binding pocket).

Differences were evident between the positions of SNPs and cancer related mutations in terms of their proximity to tree-determinants, both when calculated for each group and when accumulating data across all the kinase groups (Supporting Information Table S7b). The results shown correspond to the comparison of different types of mutations (driver, passenger, and all mutations) with all SNPs or with all nonsynonymous SNPs. Indeed, the results shown in Supporting Information Table S6 suggest that driver mutations generally behave differently to nonsynonymous SNPs, and that they are closer to important specific functional sites in each protein group (tree-determinant residues) than nsSNPs, with the exception of the AGC and TK groups. By contrast, passenger mutations cannot generally be differentiated from nonsynonymous SNPs, except in the CMGC, CK1, and TK groups, although being passenger mutations they were located differently when compared with all SNPs as a whole (with the exception of the CMGC, Other, and STE groups), further away from the specific functional regions than SNPs. These results indicate that drastic changes (represented by nonsynonymous SNPs) are more directly associated with regions implicated in specific recognition than mutations like passenger mutations, which do not seem to be particularly closely associated to regions where they may disrupt the specific network of interactions.

## Examples of well-characterized disease-associated mutations affecting kinase function

Our general analysis of the distribution of mutations in protein kinases not only provides an overview of their relationship to function and structure, but also provide an insight into their specific biomedical implications. Hence, we analyzed these relationships in three well-known scenarios.

We have reflected all the information accumulated for the protein kinase domains in a single structure used as a framework. The accumulation of information on mutations clearly increases the significance of the results and makes the interpretation of the general distribution of mutations more reliable. The assumption is that the key regions for the structure-function of the kinases are common to all of them and they can be used as a reference for the interpretation of the mutations in any of the kinases.

Mutations in the human insulin receptor (INSR) gene are known to be associated with disease,[34–37] and defects in INSR are the cause of acanthosis nigricans Type A insulin-resistant diabetes mellitus (IRAN Type A, MIM:610549). This syndrome is characterized by severe insulin resistance, including a failure to respond to exogenous insulin, which therefore increases the probability of developing endometrial cancer and other diseases.

Several studies have reported that the altered kinase activity of INSR may be fundamental in the accelerated growth in young females with this syndrome.[38] Indeed, mutated INSR exhibited diminished expression in cultured fibroblast, with the remaining receptors show impaired activity.[37] Patients with Type A insulin resistance and decreased insulin receptor kinase activity caused by a normal insulin-binding mutant, but with defects in the phosporylation mechanism, have also been described.[35,36] Recent studies have further characterized the relationship between insulin resistance and disease (For example, see Refs. 39, 40). Moreover, studies have also associated acanthosis nigricans with mutations in other kinases, such as the fibroblast growth factor receptors II and III (FGFR2_HUMAN and FGFR3_HUMAN, respectively[41–43]).

In our analysis, we mapped a mutation in the human insulin receptor (INSR_HUMAN) protein that introduced an Alanine-Threonine shift in the ATP binding pocket of the protein, A1161T (residue 173 in the model, Fig. 4). This mutation was described as a driver mutation in our analysis and it was shown to be part of the binding site in FireDB, as well as being important for the specificity shown by certain subfamilies given its relationship with subfamily specific residues (Tree-Determinants). This mutation in the ATP binding pocket may explain the difficulties to phosphorylate and the reduced enzymatic activity that leads to the development of the disease.

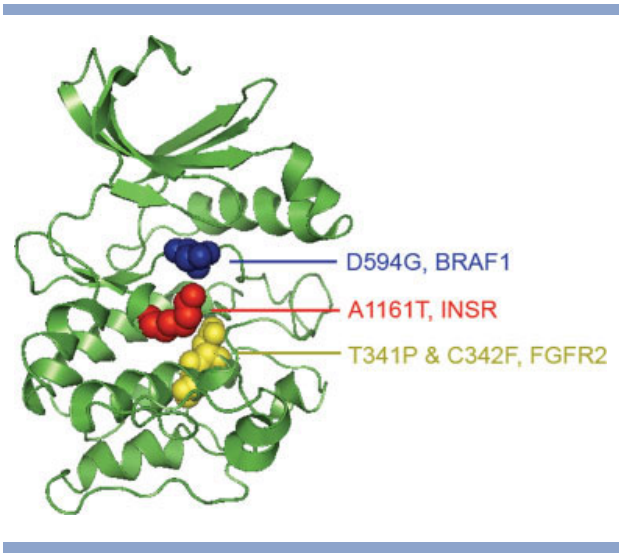



**Figure 4**
Position of several disease causing mutations in the structural model.

Another interesting example is the D594G mutation in the BRAF proto-oncogene serine/threonine protein kinase (BRAF1_HUMAN). The RAF gene family consists of three members (ARAF1, BRAF, and RAF1) that encode serine/threonine kinases, which are regulated by binding to RAS as part of the RAS-RAF-MEK-ERK-MAP kinase pathway—this plays a critical role in cell proliferation and is frequently activated in cancer cells. Previous studies showed that mutated-BRAF proteins have elevated kinase activity[44] and that the BRAF gene is somatically mutated in non-Hodgkin lymphoma, a broad group of cancers affecting the immune system. Hence, the RAS-RAF kinase pathway may be regulated by somatic mutations of BRAF in some NHLs.[45] The SAAPdb database[46] also described this mutation as a pathogenic deviation involved in the disruption of the binding site, interacting surface, quaternary structure, and the essential scaffolding of hydrogen bonds. We described D594G as a driver mutation which introduces an Asp/Gly change into the activation loop of the kinase family (position 190 in our model, see Fig. 4). Moreover, the results highlighted the importance of this amino acid in ATP binding, as described by FireDB and as a residue providing family-specific binding. These descriptions are fully coherent with the important role of this mutation in NH lymphoma development.

Finally, we considered the potential consequences of two mutations in the human fibroblast growth factor Type 2 receptor (FGFR2_HUMAN), T341P, and C342F (representing positions 233 and 234 in the structural model, respectively, see Fig. 4), to demonstrate that the analysis may also provide insights into complex diseases caused by more than a single mutation. Several diseases caused by uncontrolled cell growth have been associated with defects in FGFR2. Among them, Pfeiffer syndrome (Acrocephalosyndactyly Type V, MIM:101600) and the related Crouzon syndrome (Craniofacial dysostosis Type I, MIM:123500) have been reported to be associated with the two mutations of interest—T341P in PS and CS, C342F in CS. We found that T341P is a driver mutation introducing a Threonine-Proline change in a buried conserved position, confirming a common association of proline mutations with the disease.[5] In addition, the C342F mutation is a somatic driver mutation in a buried residue. Both mutations were considered to be related to binding specificity, indicative of the importance of these two positions for protein function.

## DISCUSSION

We have analyzed the distribution of point mutations in protein kinases with the aim of characterizing the structural and functional implications of different types of mutations focusing exclusively on the protein kinase catalytic domain, for which a representative number of structures are available. The human kinome is particularly amenable to this type of study, since it includes one of the better characterized protein superfamilies at the structural and functional level, and a large number of cancer-associated mutations have been published for these proteins.[6,7,47] We have analyzed the main characteristics of protein kinases: (a) sequence conservation as an indicator of evolutionary important regions; (b) localization in regions that are structurally conserved across the superfamily and positions characterized by their solvent accessibility; (c) organization of mutations with respect to the functional regions defined as general protein kinase activity or binding sites, and binding sites for specific effectors. Finally, we compared the distribution of point mutations in the largest protein kinase groups (i.e., AGC, CAMK, CK1, CMGC, Other, STE, TK, and TKL).

The comparison between nonsynonymous and synonymous SNPs shows that there are significant differences between the two categories in terms of sequence and structural conservation, the distance to active/binding site residues, and the distance to specific binding sites. As expected synonymous replacements are tolerated in all protein regions, whereas mutations that involve a change of amino acid (nonsynonymous) are more tolerated in regions distant from the important regions, which include evolutionary conserved residues and those involved in structural and functional features.

Driver and passenger mutations (as defined by Refs. 6, 7) also display clear differences in their distributions in protein kinases. Driver mutations are closer to regions important for protein function and structure, including ATP and substrate-binding sites, conserved residues or domains, and the apolar core of the protein, as well as residues that confer specificity to the kinase group. This

distribution is compatible with the proposed role of driver mutations in the onset of tumor formation (disease association). By contrast, passenger mutations tend to be further away from key or conserved positions, and they do not cluster around binding sites and regions that infer specificity. This distribution suggests that these mutations do not have a general disruptive effect, which is compatible with the model proposed whereby they accumulate during the process of tumor development.

The experimental information currently available for the various mutations is very limited, since the immense majority is derived from genetic studies and has not been followed by validation assays. Therefore, it is unfortunately impossible to distinguish between gain and loss of function mutants, i.e., mutations activating and deactivating the kinase activity, information that could provide very interesting insights into the molecular mechanisms associated to the mutations.

We further compared cancer-associated mutations and single-nucleotide polymorphisms by splitting the protein kinases into their different groups. Interestingly, we find significant differences in terms of conservation and accessibility, whereby SNPs tend to lie further away from conserved sequences than mutations. This finding supports the idea that SNPs are widespread in the population and have milder functional and structural consequences than mutations specifically accumulated in tumors.

Indeed, driver mutations show an even stronger tendency to occupy positions that are important to the protein when compared with SNPs (both all SNPs and nonsynonymous SNPs), including buried residues, active/binding sites, conserved, and family-specific regions. This phenomenon reinforces the idea that driver mutations have even more critical effects on protein structure/function. In addition, this strong tendency is not observed for passenger mutations that apparently affect protein function less and that are possibly more easily tolerated by cells (summarized in Supporting Information Figs. S7 and S8).

The trends observed here for cancer-associated somatic mutations are similar to the ones describe recently by Torkamani et al.[48] for inherited polymorphisms confirming the relation between disease-associated amino acid changes and key regions for protein function/structure.

## CONCLUSIONS

We have shown that the collection of cancer-associated mutations, and in particular those more likely to be related with the initial stages of tumorigenesis, are prone to be related to essential facets of protein function, including perturbation of important aspects of binding to substrates and recognition by effectors. The so-called driver mutations seem to accumulate less in key structural regions, which could be interpreted as having a reduced tendency to alter the protein globally, in contrast to passenger mutations that may produce a complete disruption of protein activity and folding when introduced in a later phase of tumor progression.

These tendencies are different to the ones of genetically inherited mutations (SNPs), where nonsynonymous SNPs associated to disease but widespread in the population, tend to avoid protein regions that could be detrimental for protein function. This tendency contrasts with the one for cancer-associated-driver-mutations, which seem to be more directly implicated in the alteration of protein function.

These results provide a better understanding of the relationship of cancer-associated mutations with protein function, and in particular a more clear definition of the implication of mutations in protein kinases for cancer progression. The modularity of the analysis carried out here and the information collected in terms of properties should also be useful for the development of better predictors of the consequences of point mutations and their implications for tumor progression.

## MATERIALS AND METHODS

### Sequences of protein kinase domains using KinBase

The KinBase resource (http://www.kinase.com/kinbase[10]) is a repository storing the currently accepted classification of eukaryotic protein kinases, which are categorized into two main groups: "conventional" protein kinases (ePKs) and "atypical" protein kinases (aPKs). The ePKs form the largest group and they have been subdivided into eight groups by sequence similarity between the catalytic domains, the presence of accessory domains, and by considering different modes of regulation. The eight ePK families defined in KinBase are as follows: the AGC group (including cyclic-nucleotide and calcium-phospholipid-dependent kinases, ribosomal S6-phosphorylating kinases, G protein-coupled kinases, and close relatives of these kinases), the CAMKs (calmodulin-regulated kinases); the CK1 group (casein kinase 1 and close relatives); the CMGC group (including cyclin-dependent kinases, mitogen-activated protein kinases, CDK-like kinases, and glycogen synthase kinase); the RGC group (receptor gyanylate cyclase kinases); the STE group (MAP Kinase cascade kinases), Tyrosine kinase group (TKs); and the TKL group (Tyrosine kinase like family), which are a cluster of serine-threonine kinases resembling TKs. Another broad, miscellaneous group called "other" is also considered for those proteins that do not fit in any of the predefined sets.

At the time of the analysis, KinBase contained 620 human protein sequences of which 516 correspond to protein kinases not considered to be pseudogenes. Although kinases described as pseudogenes are tran-

scribed and might even have a residual or scaffolding function,[10] kinase pseudogenes were not mapped onto Uniprot (SwissProt/Trembl) since many of them are partial transcripts or have stop codons in their sequence. Since KinBase does not directly map its entries onto Uniprot, this mapping was performed using a BlastP[49] search for each kinbase sequence against a custom database containing all entries in Uniprot annotated as human protein kinase domain. Once the mapping was performed, we were able to map 488 Kinbase identifiers to a valid Uniprot entry, 474 of them (97.13%) at sequence identity levels of at least 95%.

### Generation of a consensus model summarizing structural domains of protein kinases

A consensus model of the basic structure of the kinase domain was created. This consensus model represents the average structure of a large number of kinases in the human kinome and therefore, it is useful to summarize the global characteristics of these structures. To build the model, we first selected MAP3K1 as a standard representative sequence of the family from a manually curated multiple sequence alignment of the human kinome constructed using the MUSCLE alignment package.[50] The selected sequence was submitted to Modeller,[51] assembling the models created using all the closely related PDB template structures returned from a BLAST search.[49] The predicted model is represented in Figure 1 where all important functional areas are represented.

### Selection and classification of SNPs

Nowadays, the most commonly cited database for storing information on SNPs is dbSNP,[52] which currently contains several millions of validated SNPs from humans and other species. The information stored in dbSNP can also be accessed through the Ensembl[53] application programming interface (API). Every record gathered using this API contained at least a valid Swissprot identifier, the amino acid change of the mutation, the synonymous/nonsynonymous character of the polymorphism, and the position in the sequence of the protein. To get the equivalent position in each 3D structure, as recovered from the Protein Data Bank,[54] the SPICE-DAS alignment tool[55] was used to recover the sequence-structure equivalences. When applied to the 488 Kinbase protein domain identifiers, after restricting the polymorphisms to those lodging in the protein kinase domain, 569 SNPs were recovered, 263 (46.22%) of which were subclassified as nonsynonymous, whereas the remaining 306 (53.78%) were annotated as synonymous.

### Selection and classification of somatic mutations

Large-scale systematic genotyping studies have been conducted to identify mutated genes in human cancer genomes. These studies focused on tumor type (Colon and Breast)[7] or on the protein kinase superfamily,[6] and they led to the identification of more than 3500 somatic mutations. Further statistical analysis, based on mutation rates for instance, or on synonymous versus nonsynonymous mutation ratios, have been used to classify genes as "drivers" (i.e., causally involved in oncogenesis) and "passengers" (incidentally involved in oncogenesis and most likely arising during tumor development). In practical terms, where no additional evidence is present for a mutation, that driver/passenger assignment for the gene is inherited by the mutations in that gene. Those mutations along with the corresponding amino acid change, their driver/passenger character, and their sequence/structure positions (calculated using the SPICE server), were stored. The analysis performed focuses on the protein kinase catalytic domain. The protein kinase domain subset comprises 140 mutations, 73 (52%) driver and 67 (48%) passenger mutations.

### Assigning a conservation score to each position in the alignment

For each position in the alignment, conservation was measured in terms of Shannon's entropy,[56] which is a measure of the variability in the distribution of elements in a set, as described by the formula:

$$-\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \qquad (1)$$

where $p(x_i)$ is the probability of having element $x_i$ in bin i for that distribution.

Conservation was measured for each position in the alignment in two different scenarios. The most straightforward one was in a context of identity, where 21 bins were characterized, each reflecting an amino acid (plus an extra bin for gaps). In the second scenario, the bins represent different physical-chemical properties of the amino acids, which permits clustering in the context of similarity, closely related but conceptually different to the previous one. The seven groups were defined as follows: aliphatic (A, C, I, L, M, V), positive (R, K), polar (N, Q, S, T), negative (D, E), special (G,P), aromatic (H, W, Y, F), and the additional gap category.

The positions in the alignment were labeled as conserved if their Shannon's entropy was less than 0.20. Additionally, in both the identity and similarity scenarios, positions with more than 75% gaps in the corresponding multiple sequence alignments were directly labeled as not conserved.

### Calculation of conservation with AL2CO

AL2CO[28] is a program to calculate a conservation index for each position in a multiple sequence alignment using several methods. Amino acid frequencies at each

position are estimated and the conservation index is calculated for these frequencies. We used the AL2CO option to weight sequences to correct for the unequal distances between different sequence pairs in the alignment and the matrix score that gives the most common position occupied by residues with similar physico-chemical properties. Finally, we labeled as conserved those residues with a normalized conservation index threshold of 70%.

### Calculation of structural conservation

All human kinases in the data set were mapped to PDB chains. These chains were then mapped to v3.1 of CATH to assign individual domains and the majority of kinases mapped to CATH superfamilies 1.10.510.10 and 3.30.200.20 (phosphotransferase domain and phosphorylase kinase, respectively).

To calculate the structural conservation for each amino acid, it was necessary to generate a multiple structure alignment using CORA.[57] All 183 human kinase structures in the CATH superfamily were first clustered (complete-linkage) at 35% sequence identity to ensure that a sufficiently diverse set of kinases was used for the calculation and then, representatives (S35Reps) were taken from these 46 clusters. Although CORA is generally able to calculate accurate multiple structure alignments of all relatives from an entire superfamily, problems can occur in especially diverse folds where there is substantial structural variation. To assess the extent of structural diversity in the kinase superfamily, all S35Reps were aligned on a pairwise basis using the SSAP structure comparison algorithm. The S35Reps were then grouped together if they shared Simax score $< 3$ Å, producing four structurally similar groups (SSG). CORA was then used to align all SSGs.

Each fully aligned position in the SSG alignment was scored for structural conservation using a modification of the method presented in Ref. 57. Vectors between C-beta atoms were calculated for a given pair of aligned residues in a given pair of structures to all other equivalent positions (including gaps, which scored 0). The global score for each alignment position was then calculated as a sum of all the pair-wise protein scores. The score was then normalized across the whole alignment in the range 0–10, with 10 representing the most highly conserved positions. Positions with a conservation score of at least 8 were considered to be conserved.

A modification of this score was also made to focus on the local structural conservation of each position, where vectors were only calculated if they were between residues within 10 Å of one another.

### Calculation of the accessibility with Naccess

Naccess is a stand-alone program that calculates the solvent accessible area by rolling a "probe" with van der Waals radius over the surface of the molecule. We defined "buried" residues if their relative accessible surface area exposed to the "probe" is less than or equal to 16% of the total surface of the residue (also used in Consurf[58]).

### Xd analysis

To assess the significance of the proximity of different sets of mutations to specific areas of the protein (buried, functional, conserved, etc), we use the Xd measure introduced previously by Ref. 25. The most relevant characteristic of this measure of differences between distributions is that it weights more those positions lodging in bins in the proximity of the studied features instead of considering equally informative the complete distribution of distances. In other words, it gives more importance to differences in the distributions of residues close to the important regions than to differences localized in distant regions.

$$Xd = \sum_{i=1}^{i=n} \frac{P_{ic} - P_{ia}}{d_i xn} \qquad (2)$$

where $n$ is the number of distance bins in the distributions, $d_i$ is the upper limit for each bin, $P_{ic}$ is the percentage of residues with distance between $d_i$ and $d_{i-1}$, and $P_{ia}$ is the same percentage for all residues in the protein. Defined in this way, positive values of Xd indicate that the population of residues shifts to smaller distances with respect to the population of all residues. In practice, we use a difference of Xd values of 0.75 to indicate distributions of residues that are significantly different with regards their proximity to previously defined areas of the protein.

### Active site retrieval using FireDB

The FireDB database[30] contains a comprehensive curated set of substrate-binding and catalytic residues, extracted directly from PDB[54] or from the Catalytic Site Atlas.[59] FireDB binding residues for the various kinases were mapped into the general model using the corresponding multiple structure alignment.

### Prediction of tree determinant positions

S3det [Rausell et al., in preparation] is a novel implementation of the sequence space approach[31] using multiple correspondence analysis.[60] This new system is fully automated and has been optimized to detect groups of proteins within a kinase group with potential functional specificities, and to identify the residues that contain more information about that sequence classification.

S3det is based on the simultaneous vectorial representations of sequences and residues within a multiple sequence alignment (MSA) in related spaces. Once both protein and residue spaces have been optimally decomposed in their main relevant sources of variation, S3det

supplies, by means of an unsupervised clustering algorithm, an automated identification of the putative groups of proteins that will be regarded as different functional families within the MSA. After these protein families have been established, they are linked with the residue space to automatically assign the set of residues that uniquely characterizes each group. Those residues are predicted to be the functional specific determinants within the protein family. S3det can naturally exploit a supervised family classification, analogous to the MCdet method previously developed by our group.[32]

## ACKNOWLEDGMENTS

The authors want to thank David de Juan, Antonio Rausell, Anaïs Baudot, Eduardo León, Gonzalo Lopez, Ana Rojas, and Michael L. Tress for their help, interesting discussion, and ideas.

## REFERENCES

1. Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 1998;8:1229–1231.
2. Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. Genome Res 1998;8:748–754.
3. Schattner P, Diekhans M. Regions of extreme synonymous codon selection in mammalian genes. Nucleic Acids Res 2006;34:1700–1710.
4. Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM. Silent polymorphisms speak: how they affect pharmacogenomics and the treatment of cancer. Cancer Res 2007;67:9609–9612.
5. Torkamani A, Schork NJ. Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family. Genomics 2007;90:49–58.
6. Greenman C, Stephens P, Smith R, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446:153–158.
7. Wood LD, Williams Parsons D, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JKV, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PVK, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B. The genomic landscapes of human breast and colorectal cancers. Science 2007;318:1108–1113.
8. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer 2004;4:177–183.
9. Garber K. The second wave in kinase cancer drugs. Nat Biotechnol 2006;312:1175–1178.
10. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science 2001;298:1912–1914.
11. Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, et al. Targets of the cyclin-dependet kinase Cdk1. Nature 2003;425:859–864.
12. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, et al. Global analysis of protein phosphorylation in yeast. Nature 2005;438:679–684.
13. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat 2001;17:263–270.
14. Ferrer-Costa C, Orozco M, De La Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 2002;315:771–786.
15. Wang Z, Moult J. Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. Proteins 2003;53:748–757.
16. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–3814.
17. Ferrer-Costa C, Orozco M, De La Cruz X. Sequence-based prediction of pathological mutations. Proteins 2004;57:811–819.
18. Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, De La Cruz X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 2005;21:3176–3178.
19. Yue P, Moult J. Identification and analysis of deleterious human SNPs. J Mol Biol 2006;356:1263–1274.
20. Torkamani A, Schork NJ. Accurate prediction of deleterious protein kinase polymorphisms. Bioinformatics 2007;23:2918–2925.
21. Fraser H, Plotkin J. Using protein complexes to predict phenotypic effects of gene mutation. Genome Biol 2007;8:R252.
22. Bromberg Y, Yachdav G, Rost B. SNAP predicts effect of mutations on protein function. Bioinformatics 2008;24:2397–2398.
23. Mort M, Ivanov D, Cooper DN, Chuzhanova NA. A meta-analysis of nonsense mutations causing human genetic disease. Hum Mutat 2008;29:1037–1047.
24. Liu J, Zhang Y, Lei X, Zhang Z. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. Genome Biol 2008;9:R69.
25. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol 1997;271:511–523.
26. Graña O, Baker D, Maccallum RM, Meiler J, Punta M, Rost B, Tress ML, Valencia A. CASP6 assessment of contact prediction. Proteins 2005;61(Suppl 7):214–224.
27. Izarzugaza JMG, Graña O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7. Proteins 2007;69(Suppl 8):152–158.
28. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001;17:700–712.
29. Knight JD, Qian B, Baker D, Kothary R. Conservation, variability and the modeling of active protein kinases. PLoS 2007;2:e982.
30. López G, Valencia A, Tress ML. FireDB—a database of functionally important residues from proteins of known structure. Nucleic Acids Res 2007;35:D219–D223.
31. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2:171–178.
32. Pazos F, Rausell A, Valencia A. Phylogeny-independent detection of functional residues. Bioinformatics 2006;22:1440–1448.
33. Dhillon AS, Hagan S, Rath O, Kolch W. MAP kinase signaling pathways in cancer. Oncogene 2007;26:3279–3290.
34. Kahn CR, Flier JS, Bar RS, Archer JA, Gorden P, Martin MM, Roth J. The syndromes of insulin resistance and acanthosis nigricans: insulin receptor disorders in man. N Engl J Med 1976;294:739–745.
35. Grigorescu F, Flier JS, Kahn CR. Defect in insulin receptor phosphorylation in erythrocytes and fibroblasts associated with severe insulin resistance. J Biol Chem 1984;259:15003–15006.
36. Grunberger G, Comi RJ, Taylor SI, Gorden P. Tyrosine kinase activity of the insulin receptor of patients with type A extreme insulin resistance: studies with circulating mononuclear cells and cultured lymphocytes. J Clin Endocrinol Metab 1984;59:1152–1158.
37. Prince MJ, Smith FE, Peters EJ, Stuart CA. Functional characteristics of decreased insulin receptors on fibroblasts obtained from a subject with severe insulin resistance and acanthosis nigricans. Diabetes 1986;35:148–154.
38. Bromberg Y, Rost B. Comprehensive in silico mutagenesis highlights functionally important residues in proteins. Bioinformatics 2008;24:i207–i212.

39. Caceres M, Teran CG, Rodriguez S, Medina M. Prevalence of insulin resistance and its association with metabolic syndrome criteria among Bolivian children and adolescents with obesity. BMC Pediatr 2008;8:31.
40. Bonaventure J, El Ghouzzi V. Molecular and cellular bases of syndromic craniosynostoses. Expert Rev Mol Med 2003;5:1–17.
41. Leroy JG, Nuytinck L, Lambert J, Naeyaert JM, Mortier GR. Acanthosis nigricans in a child with mild osteochondrodysplasia and K650Q mutation in the FGFR3 gene. Am J Med Genet A 2007;143:3144–3149.
42. Zankl A, Elakis G, Susman RD, Inglis G, Gardener G, Buckley MF, Roscioli T. Prenatal and postnatal presentation of severe achondroplasia with developmental delay and acanthosis nigricans (SADDAN) due to the FGFR3 Lys650Met mutation. Am J Med Genet A 2008; 146:212–218.
43. Fonseca R, Costa-Lima MA, Cosentino V, Orioli IM. Second case of Beare-Stevenson syndrome with an FGFR2 Ser372Cys mutation. Am J Med Genet A 2008;146:658–660.
44. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, Hargrave D, Pritchard-Jones K, Maitland N, Chenevix-Trench G, Riggins GJ, Bigner DD, Palmieri G, Cossu A, Flanagan A, Nicholson A, Ho JW, Leung SY, Yuen ST, Weber BL, Seigler HF, Darrow TL, Paterson H, Marais R, Marshall CJ, Wooster R, Stratton MR, Futreal PA. Mutations of the BRAF gene in human cancer. Nature 2002;417:949–954.
45. Lee JW, Yoo NJ, Soung YH, Kim HS, Park WS, Kim SY, Lee JH, Park JY, Cho YG, Kim CJ, Ko YH, Kim SH, Nam SW, Lee JY, Lee SH. BRAF mutations in non-Hodgkin's lymphoma. Br J Cancer 2003;89:1958–1960.
46. Hurst JM, McMillan LEM, Porter CT, Allen J, Fakorede A, Martin ACR. The SAAPdb web resource: A large-scale structural analysis of mutant proteins. Hum Mutat 2009;30:616–624.
47. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE. The consensus coding sequences of human breast and colorectal cancers. Science 2006;314:268–274.
48. Torkamani A, Kannan N, Taylor SS, Schork NJ. Congenital disease SNPs target lineage specific structural elements in protein kinases. Proc Natl Acad Sci USA 2008;105:9011–9016.
49. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
50. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004;5:113.
51. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 2003;374:461–491.
52. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–311.
53. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E. Ensembl 2007. Nucl Acids Res 2007;35:D610–D617.
54. Berman HM, Westbrock J, Feng Z, Gillilan G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
55. Prlić A, Down TA, Hubbard TJP. Adding some SPICE to DAS. Bioinformatics 2005;21(Suppl 2):ii40–ii41.
56. Shannon CE. A mathematical theory of communication. Bell Syts Tech J 1948.
57. Orengo CA. CORA—topological fingerprints for protein structural families. Protein Sci 1999;8:699–715.
58. Glaser F, Rosenberg Y, Kessel A, Pupko T, Ben-Tal N. The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. Proteins 2005;58:610–617.
59. Porter CT, Bartlett GJ, Thornton JM. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 2004;32:D129–D133.
60. Greenacre MJ. Theory and application of correspondence analysis. London: Academic Press; 1984.

Izarzugaza et al. BMC Bioinformatics 2011, 12(Suppl 4):S1
http://www.biomedcentral.com/1471-2105/11/S4/S1

BMC Bioinformatics

RESEARCH Open Access

# Characterization of pathogenic germline mutations in human Protein Kinases

Jose MG Izarzugaza[1,2*], Lisa EM Hopcroft[2], Anja Baresic[2], Christine A Orengo[2], Andrew CR Martin[2*], Alfonso Valencia[1*]

*From* ECCB 2010 Workshop: Annotation interpretation and management of mutations (AIMM) Ghent, Belgium. 26 September 2010

## Abstract

**Background:** Protein Kinases are a superfamily of proteins involved in crucial cellular processes such as cell cycle regulation and signal transduction. Accordingly, they play an important role in cancer biology. To contribute to the study of the relation between kinases and disease we compared pathogenic mutations to neutral mutations as an extension to our previous analysis of cancer somatic mutations. First, we analyzed native and mutant proteins in terms of amino acid composition. Secondly, mutations were characterized according to their potential structural effects and finally, we assessed the location of the different classes of polymorphisms with respect to kinase-relevant positions in terms of subfamily specificity, conservation, accessibility and functional sites.

**Results:** Pathogenic Protein Kinase mutations perturb essential aspects of protein function, including disruption of substrate binding and/or effector recognition at family-specific positions. Interestingly these mutations in Protein Kinases display a tendency to avoid structurally relevant positions, what represents a significant difference with respect to the average distribution of pathogenic mutations in other protein families.

**Conclusions:** Disease-associated mutations display sound differences with respect to neutral mutations: several amino acids are specific of each mutation type, different structural properties characterize each class and the distribution of pathogenic mutations within the consensus structure of the Protein Kinase domain is substantially different to that for non-pathogenic mutations. This preferential distribution confirms previous observations about the functional and structural distribution of the controversial cancer driver and passenger somatic mutations and their use as a proxy for the study of the involvement of somatic mutations in cancer development.

## Background

Point mutations of nucleotide bases are a mechanism of crucial importance in the evolution of proteins, and hence in the evolution of organisms. A biologically relevant class of point mutation, accounting for about 90% of sequence polymorphisms [1] at an overall frequency of about one per 1000 bases [2] is the single nucleotide point mutation or PM. Traditionally, polymorphisms are classified according to their genomic location into coding or non-coding. Coding PMs can be further classified depending on whether the resulting protein product is changed owing to the genomic polymorphism. Non-synonymous PMs (nsPMs) are those that alter the amino acid sequence of the protein product through either amino acid substitution or the insertion of truncation mutations. We refer to those which generate a single amino acid substitution as 'single amino acid polymorphisms' or SAAPs. In contrast, synonymous PMs (also referred as silent or sPM) are those that do not alter the amino acid sequence of the protein product expressed. A particular case of PMs corresponds to single nucleotide polymorphisms (SNPs): those germline

* Correspondence: jmgonzalez@cnio.es; andrew@bioinf.org.uk; avalencia@cnio.es
[1]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernandez Almagro 3, E28029 Madrid, Spain
[2]Institute of Structural and Molecular Biology, Division of Biosciences, University College London, Gower Street, London WC1E 6BT, United Kingdom
Full list of author information is available at the end of the article

mutations frequently found (>1%) in normal individuals and considered neutral. A major effort to catalogue and annotate SNPs is dbSNP [3]. Although most amino acid changes are tolerated in the native protein structure, not all PMs are neutral. An increasing number of mutations are prone to be associated with aberrant phenotypes and disease. Disease-associated mutations occur at much lower frequencies in the population and have a severe effect on phenotype. Here, we use the term 'pathogenic deviation' (PD hereafter) to refer to any single base change reported to be correlated with disease. Although both PDs and nsSNPs result in a change in the expressed protein product, the former are reported to have a severe effect on phenotype whereas nsSNPs are expected to have a non-deleterious phenotypic effect.

About 1% of all human genes are known to contribute to cancer as a result of acquired mutations. The family of genes most frequently contributing to cancer is the Protein Kinase gene family [4] which is implicated in a huge number of tumorigenic functions including immune evasion, proliferation, antiapoptotic activity, metastasis and angiogenesis, possibly due to the simplicity of the mechanism of attaching an ATP-derived phosphate to a substrate protein [5]. Protein Kinases are one of the most ubiquitous families of signaling molecules in the human cell, accounting for approximately 2% of the proteins encoded by the human genome [6]. Protein Kinases show a wide-scale similarity both at sequence and structure level, attributable to the fact that all kinases transfer the terminal phosphate of ATP to a serine, threonine or tyrosine residue in a target protein. Empirical studies to date also suggest a common, with a few exceptions, catalytic mechanism whereby ATP and an active site divalent cation are bound in identical manners and phospho-transfer is carried out by a shared set of amino acids [7]. Studies [8,9] on yeast models have shown that kinases can be very promiscuous, phosphorylating a huge number of different protein substrates albeit showing remarkable specificity. This inconsistency suggests that kinases have a region committed to the general function of catalysis, with another region (or regions) customizable to confirm substrate specificity to the enzyme without any particular need to alter fold, compromise ligand binding or modify the subsequent reaction mechanism. Protein Kinases are a thoroughly studied protein family and a plethora of mutations have been previously reported in the literature [10]. These studies often include evidence of association with disease. Concomitantly, several efforts [11,12] are devoted to the prediction of the pathogenicity of somatic kinase mutations in cancer samples. These mutations are classified into two main categories: those that are involved in cancer onset and development –driver mutations– and those that are biologically neutral –passenger– mutations. For a detailed review, see Baudot *et al.*,2009 [13]. Previous work [14] characterized the preferential distribution of cancer driver kinase somatic mutations in regions of importance for protein function, including disruption of substrate binding and/or effector recognition at family-specific positions, often avoiding structurally relevant positions.

The objectives of the work presented here are twofold. Firstly, we wanted to clarify whether the trends detected for driver somatic kinase mutations can be extended to other disease related mutations independently of the nature of the mutation. Secondly, we wanted to provide additional information to the discussion on the interpretation of the role of kinase driver somatic mutations in the onset of cancer.

Consequently, we carried out a detailed multi-level comparative analysis of the differences between pathogenic and neutral (not pathogenic) germline mutations within the framework of the human kinome: amino acid composition of the polymorphisms was compared, mutations were characterized according to their potential structural effects and finally, we assessed the location of the different classes of polymorphisms with respect to kinase-relevant positions in terms of subfamily specificity, conservation, accessibility and functional sites.

## Results and discussion

### Sequence features of deleterious kinase mutations

We mapped 130 pathogenic deviations and 200 neutral SNPs to sequences within the Protein Kinase domain ($PD_{PK}$s and $SNP_{PK}$s, respectively). The native residue in the pathogenic ($PD_{PK}$) set was enriched in glycines (p=0.01) and leucines (p=0.04) when the sequences were compared using a two-sided Fisher exact test whereas it was enriched in prolines ($p=3x10^{-5}$) when the mutant amino acids were considered. As expected, in the $SNP_{PK}$ dataset none of the residue types were particularly enriched, neither in the native sequences nor in the mutated ones. Considering the native and mutant as a residue pair, three mutations were found more often in the $PD_{PK}$ dataset: leucine-proline (p=0), lysine-glutamate (p=0.02) and arginine-proline (p=0.02). Again, in the $SNP_{PK}$ dataset, no significant enrichment was found for the the wildtype-mutated pairs. The complete set of results can be found in Supplementary Table S2 in Additional file 1.

### Hypothesized structural effects of deleterious kinase mutations

SAAPdb [15] provides a characterization of the structural consequences of mutations. When these features were compared some differences between the groups appeared. $PD_{PK}$s were often observed at the interface

(p=0.03) including sites of inter-chain binding, as well as ligand binding. By contrast, $SNP_{PK}$s significantly (p=0.04) more often than $PD_{PK}$s tend to create empty spaces in the protein as a consequence of the difference in volume introduced by the side chain change, either in the core of the protein or in partially buried regions. Protein Kinase pathogenic mutations ($PD_{PK}$) compared to pathogenic mutations outside this domain ($PD_{nPK}$, standing for 'non protein kinase') produced striking results: $PD_{nPK}$s were more often explained by structural analyses: the modified residues affected stability, affected functional residues, introduced an empty region in the interior of the protein, and affected interaction prone positions (as annotated in MMDBBIND). By contrast, $PD_{PK}$ mutations were not significantly related with any of those categories. A complete description of these results can be found in Supplementary Table S3 in Additional file 1.

### Proximity of deleterious kinase mutations to kinase-specific protein features

We present here the results of mapping the different types of mutations, $PD_{PK}$s and $SNP_{PK}$s, onto a representative structural model from the Protein Kinase superfamily. The mutations were analyzed in terms of their distribution relative to evolutionary conserved positions and known functional regions. We were able to map 47 positions containing at least one of the 62 $PD_{PK}$ mutations and 27 positions with at least one of the 36 $SNP_{PK}$s to the consensus model (Figure 1) described in the Methods section and in previous studies [14]. The results of these analyses are summarized in Table 1.

#### *Proximity of kinase mutations to known functional regions*

***a) Kinase mutations and the catalytic region*** The active region of Protein Kinases includes the ATP binding site, the peptide-substrate binding sites and the catalytic loop implicated in the transference of the phosphate group. We defined the kinase-binding site as the set of residues extracted from the FireDB database [16]. This definition includes 32 residues (Figure 1) that directly contact the ATP in the binding pocket and that contains the five highly conserved residues that play a critical role in positioning ATP and stabilizing the active conformation in the catalytic mechanism [7] (see Figure 1). The distance distribution histograms (Figure 2, panels A and B) and the results in Table 1 (see also Figure S2 in Additional file 1, where the $PD_{PK}$s, $SNP_{PK}$s and catalytic residues were represented), showed a very strong tendency of $PD_{PK}$s to locate if not in catalytic residues, at least close to them. Indeed, 13 out of the 32 residues in the binding pocket were annotated as pathogenic whereas only three were annotated as neutral. Moreover, 2 out of the 5 residues described as essential

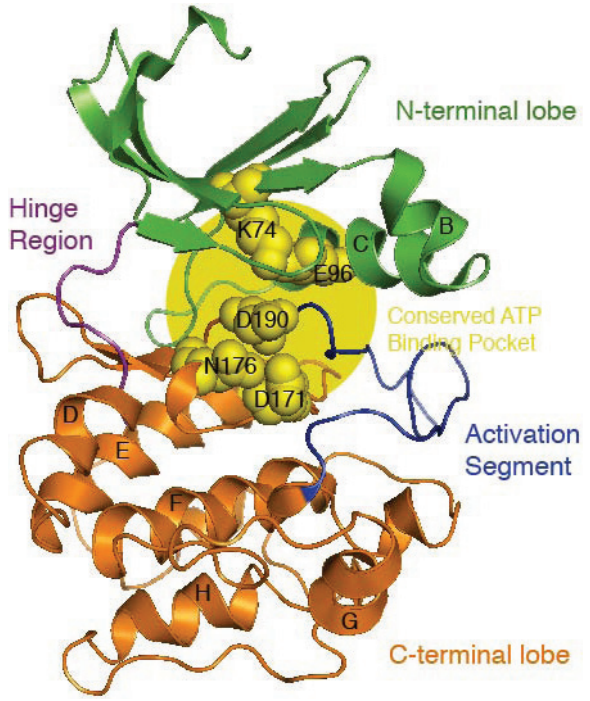



**Figure 1 Consensus model structure of summarizing the human Protein Kinase family** The model structure of human Protein Kinase, based on MAP3K1, shows the basic two-lobe kinase fold, with the N- and C-terminal (green and orange respectively) lobes joined by a hinge region (magenta). Substrate recognition is through interaction with the activation segment (blue), a region in the C-terminal lobe. The substrate-binding groove is located between the catalytic loop, the P+1 loop (activation segment), helix D, helix F, helix G and helix H. ATP binds at a site between the two lobes (yellow) that includes five conserved residues: (i) Lysine 74 that interacts with the alpha and beta phosphates of ATP and thereby stabilizing it; (ii) a nearby glutamic acid (E96) forms a salt bridge with lysine 74 increasing the stabilization network; (iii) Aspartate 171 is the catalytic base that initiates phosphotransfer by deprotonating the acceptor serine, threonine or tyrosine; (iv) Asparagine 176 interacts with a secondary divalent cation, thereby positioning the gamma-phosphate of ATP, and finally (v) Aspartate 190 which chelates the primary divalent cation, indirectly positioning ATP at the same time.

**Table 1 Results of the Xd analysis comparing $PD_{PK}$s and $SNP_{PK}$s**

| Feature | PD (Å) | SNP (Å) | ΔXd |
|---|---|---|---|
| Conservation - Shannon | 7.17 | 6.71 | -0.13 |
| Conservation - AL2CO | 8.49 | 10.49 | -0.52 |
| Structural Conservation | 7.56 | 7.00 | 2.10 |
| Accessibility - Buried | 2.94 | 3.63 | -0.83 |
| Catalytic - FireDB | 8.69 | 12.66 | -4.74 |
| Catalytic - Knight | 11.89 | 16.26 | -2.33 |
| TreeDeterminants | 6.00 | 7.94 | -1.98 |

Where PD (Å) and SNP (Å) stand for the average closest distances in Angstroms from the feature residue and the PDs and SNPs respectively, and ΔXd for the difference in Xd values. Negative values indicate that the PDs are closer to the feature residues whereas positive values indicate that SNPs are in the surroundings of feature residues.

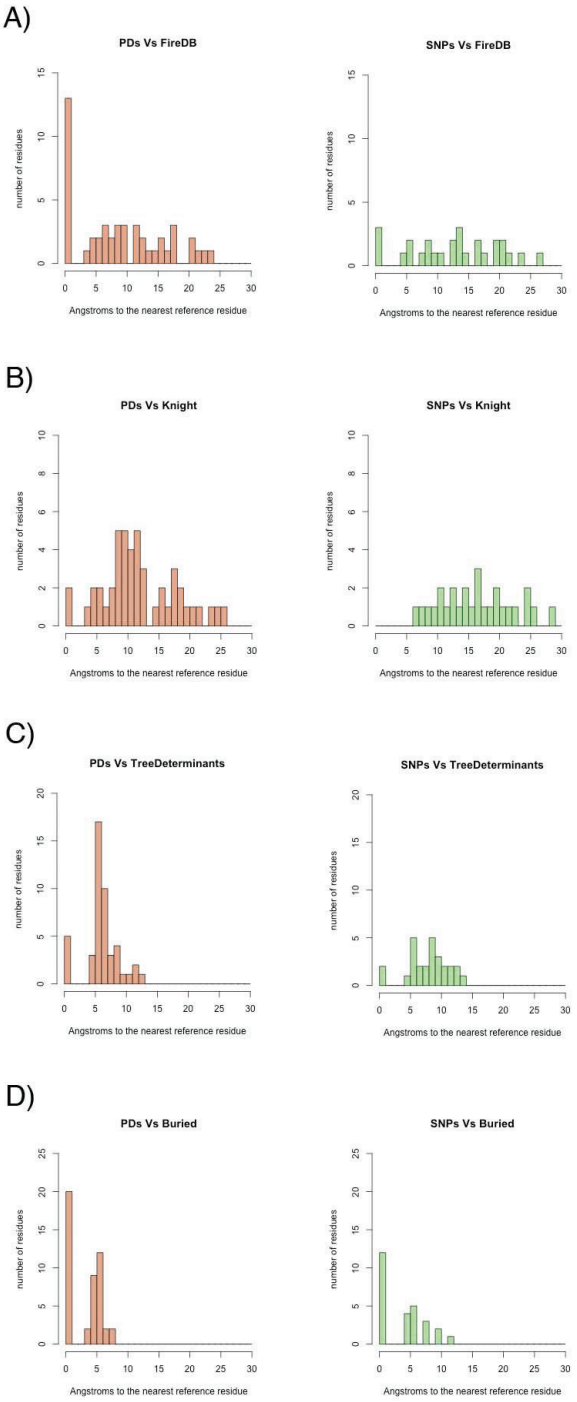



**Figure 2 Histograms of the distribution of distances between mutated resides and the analyzed features.** (A) Catalytic residues according to FireDB. (B) Catalytic Residues according to Knight et al (2007). (C) Tree-Determinants (D) Buried residues.

for the correct functioning of the ATP binding pocket [7] were annotated as $PD_{PK}$s.

***b) Kinase mutations and regions of functional subspecificity*** In this study we used the position of the tree-determinant residues as a proxy for functionally important regions in Protein Kinases, particularly those related with the specific functions of each one of the subfamilies. Residues specific to the various subfamilies of Protein Kinases were identified for each of the eight subfamilies in which KinBase categorizes the human kinome [6]. Our recent implementation of the sequence-space approach, S3det [17] identified 32 unique positions in the model as containing information relevant for differentiating between subfamilies; that is, residues that tend to be conserved in the specific subfamilies and vary to different degrees in the others (Fig. S3 in Additional file 1 depicts the distribution of the mutations along with these tree-determinant residues). Out of the 32 tree-determinants in the model, five were annotated as pathogenic (residues 50, 173, 190, 217 and 233 in the generated structural model) whereas only two were annotated as neutral. Pathogenic deviations, if not exactly in positions that were disease-associated, clustered around tree-determinant residues in general. This tendency was especially relevant for tree-determinants in the ATP binding pocket, but also appreciable in the other function specific tree-determinants. $PD_{PK}$s were closer to tree-determinant positions than neutral $SNP_{PK}$s (Fig. 2C) This was also clear in the difference of Xd values (-1.98) indicating the existence of significant differences between $PD_{PK}$s and $SNP_{PK}$s with respect to proximity to positions characterized as important for the function and subspecificity of the kinases.

***Proximity of kinase mutations to the protein core***

With the accessibility parameters defined in Methods, 99 residues were classified as buried in the kinase structural model; 20 of these buried residues are annotated as $PD_{PK}$s and 12 were annotated as $SNP_{PK}$s (three residues – 58, 197 and 218 – are described in both datasets) (Fig. S4 in Additional file 1). The distribution of distances (Fig. 2D) manifests a clear tendency of $PD_{PK}$s to be closer to buried residues. As a matter of fact, the analysis of the mean distances of mutated positions to buried residues (2.94Å and 3.63Å respectively) revealed a tendency for the pathogenic deviations ($PD_{PK}$s) to be closer to buried residues than the neutral polymorphisms. This fact was supported by a Xd difference of -0.83.

## Examples of well-characterized disease-associated mutations affecting kinase function

The statistically significant results provided after analyzing the sequence features of the single amino acid polymorphisms (see sequence features of deleterious kinase mutations section) are supported by several examples in the literature. Karkkainen *et al.*, (2000) [18] associated human hereditary lymphedema, a particular case of lymphatic obstruction, with mutations in the vascular

endothelial growth factor receptor 3 (VEGFR-3). The endothelial cells lining blood and lymphatic vessels depend on signal transduction mediated by specific receptor tyrosine kinases for their differentiation into a primary vascular plexus (vasculogenesis) and for the sprouting and splitting of new capillaries from previously existing vessels (angiogenesis). Two mutations were reported as the cause of the disease, due to the loss of tyrosine kinase activity and hence impaired downstream signaling, and a slower rate of internalization of the receptor. The suggested model highlights a mutation in the second arginine in the highly conserved HRDLAARN motif in the catalytic site to proline that facilitates the hydrogen bond between aspartate and a hydroxyl group from ATP at the binding pocket. This aspartate is critical for protein function as it is believed to act as the catalytic base in the phosphotransfer reaction. Moreover, their study revealed that a mutation from leucine to proline disturbed protein function due to both a disturbance in protein structural integrity and an interference of ATP binding. Since the pathogenic residue is located in the middle of a $\beta$-strand, introducing a restricted mobility residue such as proline disrupts the interaction with the surrounding $\beta$-strands, destabilizing the protein fold in the region. In addition, the sidechain of this leucine is part of the adenine-binding pocket, therefore alterations in such a relevant position modify the shape of the cleft and might interfere with ATP binding. With respect to the significant lysine-glutamate mutation, Mao *et al.*, (2001) [19] described the commonly cited mutation K430E (amongst others) in Bruton's tyrosine kinase (BTK, Q06187), speculating with regard to how it might cause severe XLA (X-linked agammaglobulinemia), and providing a solved crystal structure. BTK is expressed on the surface of B cells and its kinase activity is crucial in proliferation and differentiation to mature B lymphocytes. According to their mechanism, upon the trans-phosphorylation of Y551, the highly conserved K430 facilitates a hydrogen bond with E445, causing a shift of the αC helix. The K430E mutation disables the αC helix repositioning, which is crucial for the catalytic activity of BTK, hence impairing the creation of mature B cells.

### Assessing the possible functional role of relevant kinase mutations by their sequence-structure characteristics

Our general analysis of the distribution of mutations in Protein Kinases not only provided an overview of their relation with function and structure, but also provided an insight into their specific biomedical implications. In the work presented here, we summarized all the knowledge accumulated for the Protein Kinase domain in a single framework structure (Fig. 1) under the assumption that regions important for the structure/function of the kinases are common to the whole family and hence they can be used as a reference for the interpretation of the mutations in any of the individual kinases. The accumulation of information clearly increases the significance of the results provided and makes the distribution of the polymorphisms more reliable and accurate. For instance, mutations in the insulin receptor gene in humans (INSR) have been reported as disease-associated in the literature. Defects in INSR are the cause of insulin-resistant diabetes mellitus with acanthosis nigricans type A (IRAN type A, MIM:610549), a syndrome characterized by the association of severe insulin resistance manifested by marked hyperinsulinemia and a failure to respond to exogenous insulin with the skin lesion acanthosis nigricans and overian hyperandrogenism in adolescent female patients. The relation of the disease with mutations in kinases is thoroughly reported in OMIM. Recent studies have further characterized the relationship between insulin resistance and disease (for example, [20]). Moreover, several studies have associated acanthosis nigricans with mutations in other kinases, such as the fibroblast growth factor receptors II and III [21-23]. Here, we identified A1161T (residue 173 in the model), a well-characterized mutation that introduces an alanine-threonine shift in the ATP binding pocket of INSR. This mutation has been defined in our analysis as a pathogenic deviation ($PD_{PK}$). Concomitantly, it is a catalytic residue in FireDB and is important for family specificity. This perturbation of the ATP binding pocket might explain the unpaired phosphorylation and therefore the reduced enzymatic activity leading to the aberrant phenotype.

Finally, we considered not only a single mutation but a pair of consecutive mutations: T341P and C342F (positions 233 and 234 in the structural model respectively), in the human fibroblast growth factor receptor type 2 (FGFR2) to demonstrate that although only T341P is reported to be a pathogenic deviation in our dataset, the analysis can provide insights into complex diseases caused by more than one mutation at the same time. Several diseases caused by uncontrolled cell growth have been associated to defects on FGFR2. Among them, two related syndromes – Pfeiffer syndrome (PS) and Crouzon syndrome (CS) – have been reported to be associated to the pair of mutations of interest (T341P in PS and CS, C342F in CS). In our analysis, we described T341P as a mutation that introduces a change from threonine to proline in a buried sequence-conserved position. Other authors [15,24] have shown that mutations to proline are very often associated with disease. Additionally, we have characterized mutation C342F as a replacement of a buried residue. In addition, we identified both mutations as tree-determinant residues, thus considered related to binding

specificity and indicative of the importance of these two positions for protein function.

## Conclusions

We have analyzed point mutations in the structure of Protein Kinases in order to characterize the structural and functional singularities of pathogenic and neutral mutations. Although the definition of the groups is by no means stable and the groups are constantly being redefined as new studies on the pathogenicity of mutations arise, this might be used as a proxy to deepen the knowledge on the underlying mechanisms of disease. The human kinome is particularly amenable for this type of study since much is known about the structure and function of this protein family and very relevant cancer-associated mutations have been published for these proteins [11,12,14].

To address this point, pathogenic deviations mapped to the kinase domain ($PD_{PK}$) and single nucleotide polymorphisms mapped to the kinase domain ($SNP_{PK}$) were compared on the basis of sequence, hypothesized structural effect and proximity to known kinase-specific features within the framework of a modeled consensus structure, representative of the whole superfamily. At the sequence level, several mutations were differentially observed in the $PD_{PK}$ and $SNP_{PK}$ datasets: the leucine-proline mutation emerged as an interesting feature of the $PD_{PK}$ dataset: it was identified both when analyzing the native and mutant residues separately, and when analyzing the mutation pairs, and it has been identified as being indicative of disease elsewhere, both specifically in a kinase disease dataset [24] and in wider, non-disease-specific datasets [15,25,26]. In addition, replacing an arginine with a proline was more often observed in the $PD_{PK}$ dataset. Again, this mutation has been described as being associated both with diseases predicted to be related to mutations in kinases [24] and across all other diseases [15,26]. Finally, replacing the positive charge of Lysine with the bulkier, negatively charged glutamic acid sidechain is identified in the $PD_{PK}$ dataset. Unlike the well-characterized previous mutations, this is a novel observation. These finding provide evidence about the existence of distinctive differences between the two types of mutations even at such a coarse grain level as the amino acid composition.

In the second part of the analysis we focused on the singularities of the mutations in structurally and functionally relevant regions. Thus, we characterized $PD_{PK}$ mutations in terms of their structural consequences by comparing them to $SNP_{PK}$ mutations. $PD_{PK}$ mutations were more likely to occur at the interface with ligands and other protein chains and $SNP_{PK}$ mutations were more likely to introduce a cavity in the protein core. This is coherent with the previous publications concluding that disease-associated mutations in kinases often affect the site of ATP binding [16,18,19]. In addition, $PD_{PK}$ mutations were compared to other PD mutations ($PD_{nPK}$) to comment on the mechanisms by which mutations in kinases might lead to disease. In general, it is easier to explain $PD_{nPK}$s in structural terms. Most noticeably, very few $PD_{PK}$ mutations resulted in a significant cavity, an empty space, or affected protein stability at all. An increasing body of literature attributes the pathogenic nature of PDs to their destabilizing effect on native protein structure [15,27-30]. The results here, specific to the mutations in the Protein Kinase domain, contradict this trend, indicating that the pathogenicity of $PD_{PK}$s could not be simply attributed to a decreased stability of the resulting proteins. This might be explained by the fact that Protein Kinases are known to be a highly structurally conserved superfamily, while varying significantly with respect to sequence; as such, the structures must be tolerant to sequence variation. Hence, considerable flexibility must exist within the protein structure to be robust to sequence diversity. It is noticeable the relatively small number of structures available, we are aware that the results will benefit from a dedicated modelling pipeline, especially for very divergent kinases differing from the canonical PK domain. By contrast, a customized model for every single protein would not only increment the complexity of the comparative analysis of distances but would also add some uncertainties derived from the use of protein models.

In order to provide more detailed insight into the distinctive implication of pathogenic mutations in the disruption of protein function, we characterized the differences between $PD_{PK}$s and $SNP_{PK}$s by their association to kinase-specific functional and structural features. There were significant differences between the two types of mutations in terms of conservation, accessibility, distance to active/binding site residues and distance to family specific binding sites. It is interesting to compare these results with the ones previously obtained for driver/passenger mutations in Protein Kinases [14] from the works by [11,12] that constitute a subset of the many mutations stored in COSMIC [31]. Drivers are those mutations predicted to be involved in cancer onset whereas passenger mutations are those that are supposed to be accumulated during cancer progression being neutral respect to the origin of the cancer. In the previous work, the main conclusion was that driver mutations tended to be located near to important residues such as sequence conserved positions, family specific regions and active/catalytic sites whereas passenger mutations were located closer to structurally conserved regions. The difference observed here for the set of $PD_{PK}$s and $SNP_{PK}$s mimicked the one previously observed for driver and passenger mutations. Both datasets proved to be non-redundant. Only 4 residues were common to both

driver and pathogenic datasets, all of them in the human B-raf proto-oncogene. The small overlap encountered is consistent with the very different nature –germline and somatic– of the mutations in each of the disease-prone datasets. Thus, the new results presented here can be seen as a confirmation of the functional/structural role of the mutations that are more likely to be pathogenic (drivers and PDs). Given that the definition of mutations as drivers and passengers is somehow controversial (For a review, see [13]), it is good to see that the current results could be interpreted as an indirect support to the categorization of mutations into drivers (disease-associated) and passengers (disease-neutral) and their use as a proxy for the study of the involvement of somatic mutations in cancer biology.

In summary, we have confirmed that the pathogenicity of mutations within the Protein Kinase superfamily is related to essential aspects of protein function, including perturbation of substrate binding and recognition by effectors at family-specific positions. As a matter of fact, pathogenic deviations ($PD_{PK}$) accumulate in key functional regions whereas they seem to be absent from structurally relevant positions. These observations reinforce the idea that the pathogenicity of disease-associated mutations can be attributed to a disruption of native protein function while avoiding drastic changes prone to disrupt the protein globally. This tendency was not observed for neutral polymorphisms, which are apparently less disruptive of protein function and tend to be tolerated. Consequently, they are more often found in normal individuals. However, it is clear that further classification of the mutations in more specific subgroups will be necessary to provide a deeper knowledge on the mechanisms leading to disease. A typical example would be the characterization of mutations into a gain-of-function/loss-of-function on a large-scale.

The analysis presented here provides not only a characterization of the mutations, but also in some cases additional insight into the specific biomedical implications of the mutations. This type of approach will be particularly useful as part of the bioinformatics platform developed for the International Cancer Genome Consortium and other cancer genome projects. The results discussed in this work are biologically sound not only because they contribute to our understanding of the role of mutations in disease, but also because they can bee seen as a necessary step towards the development of predictors of pathogenicity based on the combination of the features analyzed here. Indeed, further development of the idea presented here might naturally derive in a classifier that making use of machine learning techniques would be able to predict the pathogenicity of any novel kinase mutation.

Moreover, although out of the scope of this work, the methodology used to analyze the distribution of the mutations in relation to a set of features can be applied to other families. However, we see (as other authors in the field [24]) advantages to the in-depth exploration of a family with more information about sequence, structure and mutations and for what it is feasible to obtain phylogenetic information to calculate subfamily specific sequence features. It remains interesting to analyze other protein families at a genome wide scale to corroborate to what extend the results shown in this work can be generalized.

## Methods

### Protein Kinase domain sequences

The KinBase resource [6] is a repository of the currently accepted classification of eukaryotic Protein Kinases. At the moment of the analysis, KinBase contained 620 human protein sequences of which 516 were Protein Kinases not defined as pseudogenes by the database curators. Although it has been described that some kinase pseudogenes are transcribed and even might have a residual or scaffolding function [32] kinase pseudogenes were not considered in the analysis performed here. KinBase does not map directly onto UniProtKB [33]. The mapping was performed using a BlastP [34] for each sequence against a custom database containing all entries in UniProtKB annotated as Protein Kinase domain for human. We were able to map 488 KinBase identifiers to a valid UniProtKB entry, 474 of them (97.13%) at sequence identity levels of at least 95%.

### Classification of the mutations

SAAPdb [15] is a database of single amino acid polymorphisms (SAAPs) mapped to protein structure. SAAPdb aims to provide likely structural effects of mutations and identify differences in potential structural consequences between neutral and pathogenic mutations. The disease dataset is derived mostly from OMIM [35] whereas the neutral dataset comes from dbSNP. For the Protein Kinase domain of the 488 kinases in SAAPdb contains 130 pathogenic deviations ($PD_{PK}$s) and 200 neutral polymorphisms ($SNP_{PK}$s). Of these 130 $PD_{PK}$s, 62 were successfully mapped to a residue in the solved PDB structure. Similarly, of the 200 $SNP_{PK}$s mapped to sequence, 36 were mapped to a PDB structure (See Supplementary Table S1 in Additional file 1). In order to create control datasets 9263 non-kinase PDs ($PD_{nPK}$) were retrieved from SAAPdb, out of them 4652 mapped to a structure. All three datasets contain only unique, non-synonymous (missense) mutations. We excluded nonsense and synonymous mutations because these have a known, truncating effect or no effect on the protein structure. A unique mutation was defined by the combination of four parameters: UniProtKB accession number and sequence position, native amino acid and mutated amino acid.

### Comparing mutations with respect to the native and/or mutant residues

To compare the mutations with respect to their sequence features, the native and mutant residues were extracted from SAAPdb, as described above. Two-sided Fisher exact tests were carried out since they allow robust comparison of datasets of disparate sizes and evaluation of contingency tables with empty cells.

### Structural effects of mutations

The SAAPdb hypothesized structural effects are fully described elsewhere [15] and are summarised briefly below. (a) Mutations affecting stability: mutations on the surface of proteins that replace a hydrophilic residue with a hydrophobic residue are identified as introducing *unfavourable hydrophobicity on the surface*. Similarly, mutations in the core that replace a hydrophobic residue with a hydrophilic residue are identified as *introducing unfavourable hydrophobicity in the core*. Buried mutations that create a *charge shift* are also identified. Using a geometric analysis of PDB structures, SAAPdb identifies mutations that affect potential *disulphide bonds*. Mutations that introduce large *cavities* in the protein core or that break *hydrogen bonds* are identified as well. (b) Mutations affecting folding: unfavourable (with respect to torsion angles) mutations from *cis-proline*, *from glycine* and *to proline* are identified. Mutations that will *clash with existing residues* are identified. (c) Mutations to UniprotKB annotated residues: mutations at the site of residues annotated by UniprotKB as functionally relevant. (d) Mutations to binding residues: PDB structures are analysed to identify residues *binding* to proteins, DNA or small molecules. These data are augmented by data from *MMDBBIND*. (e) Mutations disrupting the quaternary structure (f) Mutations to sequence conserved residues. In addition, we provide results for the summary analyses categories of *structurally explained* (where at least one structural explanation is true, i.e., any explanation apart from the sequence conservation) and *explained* (where at least one explanation is true). In order to compare the mutations with respect to their structural effects, a binary explanatory vector is calculated for each mutation and a two-sided Fisher exact test was carried out.

### Generation of a consensus model summarizing Protein Kinase structures

A consensus model (Fig.1) of the basic structure of the kinase domain was created. This consensus model represents the average structure of a large number of kinases widespread along the human kinome, and therefore it is useful to summarize global characteristics of the structures. To build the model we first selected MAP3K1 as a standard representative sequence of the family from a manually curated multiple sequence alignment of the human kinome constructed using the alignment package MUSCLE [36]. The selected sequence was submitted to Modeller [37] assembling the models created using all those closely related template PDBs structures returned from a BLAST search. The predicted model has previously been used as a consensus of the Protein Kinase domain [14]. Finally, mutations were transferred from their own PDB coordinates to the consensus model for comparison.

### Calculation of important regions

#### *Calculation of accessibility*

NACCESS (Hubbard, *unpublished*) is a stand-alone program that calculates accessible areas by rolling a probe with van der Waals radius over the surface of the molecule. A residue is defined as buried if 16% or less of the residue's surface is exposed to the probe. This is a common threshold [38] that ensures a reasonable number of buried residues.

#### *Definition of catalytic sites*

The FireDB database [16] contains a comprehensive curated set of substrate binding and catalytic residues, extracted directly from the PDB [39] or from the Catalytic Site Atlas [40]. FireDB binding residues for the various kinases were mapped into the general model using the corresponding multiple structure alignments.

#### *Prediction of Tree Determinant positions*

S3det [17] is an algorithm for the detection of groups of proteins within a family with potential functional specificities and to identify the residues that are characteristic of that group. S3det is based on the simultaneous quantitative analysis of sequences and residues within a multiple sequence alignment on related multidimensional spaces. Those residues associated with specific sequence subfamilies tend to be determinants of functional specificity and are located in functional regions of protein families, including substrate binding sites, functional sites and protein interaction sites.

### Xd analysis

To assess the significance of the proximity of different sets of mutations to areas of the protein (buried, functional, conserved, etc) we used the harmonic deviation, Xd, measure introduced previously [41].

$$Xd = \sum_{i=1}^{i=n} \frac{P_{ic} - P_{ia}}{d_i \cdot n} \quad (1)$$

Where n is the number of distance bins in the distributions, $d_i$ is the upper limit for each bin, $P_{ic}$ is the percentage of residues with distance between $d_i$ and $d_{i-1}$ and $P_{ia}$ is the same percentage for all residues in the protein. Defined this way, positive values of Xd indicate

that the population of residues is shifted to smaller distances with respect to the population of all residues. In practice we used a difference of Xd values of 0.75 to indicate distributions of residues that are significantly different regarding their proximity to previously defined areas of the protein. This threshold – albeit arbitrary – is based on manual inspection of previous results and has been proved valid in a similar context [14].

## Additional material

**Additional File 1:**

**Acknowledgements**
The authors want to thank D. Juan, A. Rausell, E. Leon, A. Carro, G. Lopez, O. Redfern and M. L. Tress for their help, interesting discussion and ideas. This work was supported by Grant BIO2007-66855 from the Spanish Ministerio de Ciencia e Innovacion. This work was supported by Grant BIO2007-66855 from the Spanish Ministerio de Ciencia e Innovacion, by grant COMBIOMED (RD07/0067/0014) from the ISCIII and by the CONSOLIDER E-science grant (CSD2007-00050).
This article has been published as part of *BMC Bioinformatics* Volume 12 Supplement 4, 2011: Proceedings of the European Conference on Computational Biology (ECCB) 2010 Workshop: Annotation, interpretation and management of mutation (AIMM). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/11?issue=S4.

**Author details**
[1]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernandez Almagro 3, E28029 Madrid, Spain. [2]Institute of Structural and Molecular Biology, Division of Biosciences, University College London, Gower Street, London WC1E 6BT, United Kingdom.

**Authors contributions**
AV, CO and AM conceived the idea. All authors planned the analysis. JMGI, LEMH and AB generated the datasets. JMGI and LEMH performed the analysis. All authors discussed the results. JMGI, LEMH and AV wrote the manuscript. All authors read and approved the manuscript.

**Competing interests**
The authors declare that they have no competing interests.

Published: 5 July 2011

**References**

1. Collins FS, Brooks LD, Chakravarti A: **A DNA polymorphism discovery resource for research on human genetic variation.** *Genome Res* 1998, **8(12)**:1229-31.
2. Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY: **Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms.** *Genome Res* 1998, **8(7)**:748-54.
3. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29**:308-11.
4. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4(3)**:177-83.
5. Garber K: **The second wave in kinase cancer drugs.** *Nat Biotechnol* 2006, **24(2)**:127-30.
6. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S: **The protein kinase complement of the human genome.** *Science* 2001, **298(5600)**:1912-1934.
7. Knight JDR, Qian B, Baker D, Kothary R: **Conservation, variability and the modeling of active protein kinases.** *PLoS ONE* 2007, **2(10)**:e982.
8. Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, Shah K, Shokat KM, Morgan DO: **Targets of the cyclin-dependent kinase Cdk1.** *Nature* 2003, **425(6960)**:859-64.
9. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJR, Stern DF, Virgilio CD, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF, Snyder M: **Global analysis of protein phosphorylation in yeast.** *Nature* 2005, **438(7068)**:679-84.
10. Krallinger M, Izarzugaza JMG, Rodriguez-Penagos C, Valencia A: **Extraction of human kinase mutations from literature, databases and genotyping studies.** *BMC Bioinformatics* 2009, **10(Suppl 8)**:S1.
11. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446(7132)**:153-8.
12. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber TD, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JKV, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PVK, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B: **The genomic landscapes of human breast and colorectal cancers.** *Science* 2007, **318(5853)**:1108-13.
13. Baudot A, Real F, Izarzugaza J, Valencia A: **From cancer genomes to cancer models: bridging the gaps.** *EMBO Rep* 2009, **10**:359-366.
14. Izarzugaza J, Redfern O, Orengo C, Valencia A: **Cancer-associated mutations are preferentially distributed in protein kinase functional sites.** *Proteins* 2009, **77**:892-903.
15. Hurst J, McMillan L, Porter C, Allen J, Fakorede A, Martin A: **The SAAPdb web resource: A large-scale structural analysis of mutant proteins.** *Hum Mutat* 2009, **30**:616-624.
16. López G, Valencia A, Tress ML: **FireDB–a database of functionally important residues from proteins of known structure.** *Nucleic Acids Res* 2007, **35(Database issue)**:D219-23.
17. Rausell A, Juan D, Pazos F, Valencia A: **Protein interactions and ligand binding: from protein subfamilies to functional specificity.** *Proc Natl Acad Sci USA* 2010, **107(5)**:1995-2000.
18. Karkkainen MJ, Ferrell RE, Lawrence EC, Kimak MA, Levinson KL, McTigue MA, Alitalo K, Finegold DN: **Missense mutations interfere with VEGFR-3 signalling in primary lymphoedema.** *Nat Genet* 2000, **25(2)**:153-9.
19. Mao C, Zhou M, Uckun FM: **Crystal structure of Bruton's tyrosine kinase domain suggests a novel pathway for activation and provides insights into the molecular basis of X-linked agammaglobulinemia.** *J Biol Chem* 2001, **276(44)**:41435-43.
20. Caceres M, Teran CG, Rodriguez S, Medina M: **Prevalence of insulin resistance and its association with metabolic syndrome criteria among Bolivian children and adolescents with obesity.** *BMC Pediatr* 2008, **8**:31.
21. Leroy JG, Nuytinck L, Lambert J, Naeyaert JM, Mortier GR: **Acanthosis nigricans in a child with mild osteochondrodysplasia and K650Q mutation in the FGFR3 gene.** *Am J Med Genet A* 2007, **143A(24)**:3144-9.
22. Zankl A, Elakis G, Susman RD, Inglis G, Gardener G, Buckley MF, Roscioli T: **Prenatal and postnatal presentation of severe achondroplasia with developmental delay and acanthosis nigricans (SADDAN) due to the FGFR3 Lys650Met mutation.** *Am J Med Genet A* 2008, **146A(2)**:212-8.
23. Fonseca R, Costa-Lima MA, Cosentino V, Orioli IM: **Second case of Beare-Stevenson syndrome with an FGFR2 Ser372Cys mutation.** *Am J Med Genet A* 2008, **146A(5)**:658-60.
24. Torkamani A, Schork NJ: **Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family.** *Genomics* 2007, **90**:49-58.

25. Krishnan VG, Westhead DR: **A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function.** *Bioinformatics* 2003, **19(17)**:2199-209.
26. Vitkup D, Sander C, Church GM: **The amino-acid mutational spectrum of human genetic disease.** *Genome Biol* 2003, **4(11)**:R72.
27. Wang Z, Moult J: **SNPs, protein structure, and disease.** *Hum Mutat* 2001, **17(4)**:263-70.
28. Ferrer-Costa C, Orozco M, de la Cruz X: **Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties.** *J Mol Biol* 2002, **315(4)**:771-86.
29. Ferrer-Costa C, Orozco M, de la Cruz X: **Sequence-based prediction of pathological mutations.** *Proteins* 2004, **57(4)**:811-9.
30. Yue P, Li Z, Moult J: **Loss of protein structure stability as a major causative factor in monogenic disease.** *J Mol Biol* 2005, **353(2)**:459-73.
31. Forbes S, Clements J, Dawson E, Bamford S, Webb T, Dogan A, Flanagan A, Teague J, Wooster R, Futreal PA, Stratton MR: **COSMIC 2005.** *Br J Cancer* 2006, **94(2)**:318-22.
32. Manning G, Plowman GD, Hunter T, Sudarsanam S: **Evolution of protein kinase signaling from yeast to man.** *Trends Biochem Sci* 2002, **27(10)**:514-20.
33. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E: **Infrastructure for the life sciences: design and implementation of the UniProt website.** *BMC Bioinformatics* 2009, **10**:136.
34. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-402.
35. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33(Database issue)**: D514-7.
36. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.
37. Fiser A, Sali A: **Modeller: generation and refinement of homology-based protein structure models.** *Meth Enzymol* 2003, **374**:461-91.
38. Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML: **Progress and challenges in predicting protein-protein interaction sites.** *Brief Bioinformatics* 2009, **10(3)**:233-46.
39. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235-42.
40. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, **32(Database issue)**:D129-33.
41. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A: **Correlated mutations contain information about protein-protein interaction.** *J Mol Biol* 1997, **271(4)**:511-23.

doi:10.1186/1471-2105-12-S4-S1
**Cite this article as:** Izarzugaza *et al.*: Characterization of pathogenic germline mutations in human Protein Kinases. *BMC Bioinformatics* 2011 **12**(Suppl 4):S1.